

DOCUMENT RETRIEVING DEVICE

Publication number: JP10049549

Publication date: 1998-02-20

Inventor: INABA MITSUAKI; NOGUCHI NAOHIKO; SUGANO YUJI; SATO MITSUHIRO; NOMOTO MASAKO; YASUKAWA HIDEKI

Applicant: MATSUSHITA ELECTRIC IND CO LTD

Classification:

- international: **G06F17/30; G06F17/30; (IPC1-7): G06F17/30**

- European: G06F17/30T2P4V

Application number: JP19970087328 19970324

Priority number(s): JP19970087328 19970324; JP19960156418 19960529

Also published as:

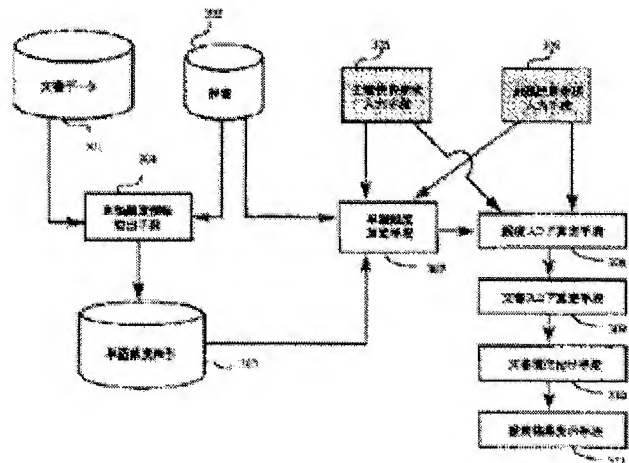
EP0810535 (A2)
US6154737 (A1)
EP0810535 (A3)
CN1172994 (A)
EP0810535 (B1)

more >>

[Report a data error here](#)

Abstract of JP10049549

PROBLEM TO BE SOLVED: To provide a document retrieving device capable of shortening the overall retrieving time including highly accurate retrieval and convergence. **SOLUTION:** A frequency score calculating means 308 calculates a frequency score indicating a matching degree between a document and a retrieval request by word frequency from the total number of documents, the number of documents in which a certain word appears, the appearance frequency of the word in each document, and the weighting parameter of the word which are outputted from a word frequency calculating means 307 and a document score calculating means 309 calculates a document score indicating a matching degree between the document and the retrieving request from the frequency score and orders the score, so that a retrieving result more close to a user's retrieving intension can be obtained.



Data supplied from the **esp@cenet** database - Worldwide

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-49549

(43) 公開日 平成10年(1998) 2月20日

(51) Int.Cl.⁵
G 0 6 F 17/30

識別記号 庁内整理番号

F I
G 0 6 F 15/403
15/40

技術表示箇所
3 4 0 B
3 7 0 A

審査請求 未請求 請求項の数7 F D (全 27 頁)

(21) 出願番号 特願平9-87328

(22) 出願日 平成9年(1997) 3月24日

(31) 優先権主張番号 特願平8-156418

(32) 優先日 平8(1996) 5月29日

(33) 優先権主張国 日本 (J P)

(71) 出願人 000005821

松下電器産業株式会社

大阪府門真市大字門真1006番地

(72) 発明者 稲葉 光昭

大阪府門真市大字門真1006番地 松下電器
産業株式会社内

(72) 発明者 野口 直彦

大阪府門真市大字門真1006番地 松下電器
産業株式会社内

(72) 発明者 菅野 祐司

大阪府門真市大字門真1006番地 松下電器
産業株式会社内

(74) 代理人 弁理士 役 昌明 (外3名)

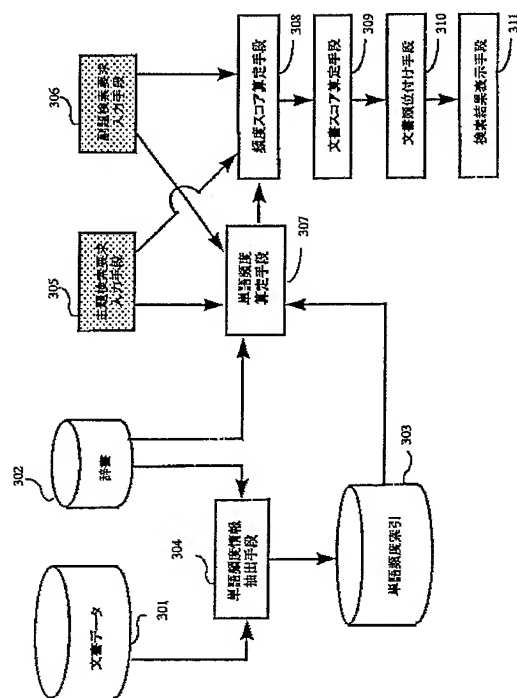
最終頁に続く

(54) 【発明の名称】 文書検索装置

(57) 【要約】

【課題】 文書データからユーザの入力した検索要求に合致する文書を探索し、その合致する度合によって順位付けを行なう文書検索装置に関するもので、従来の単語頻度のみによる文書の順位付けのもっていたユーザの検索意図に反した記事が上位に来てしまう問題点を解決し、高精度な検索と絞り込を含めた総合的な検索時間の短縮を可能にする文書検索装置の提供を目的とする。

【解決手段】 頻度スコア算定手段308は単語頻度算定手段307が出力した総文書数、単語の出現する文書数、文書における単語の出現頻度、単語の重み付けパラメータから、単語頻度による文書と検索要求の合致度合を示す頻度スコアを算出し、文書スコア算定手段309が上記頻度スコアから文書と検索要求の合致度合を示す文書スコア算出し、順位付けを行なうことによって、ユーザの検索意図により近い検索結果を得ることが可能となる。



【特許請求の範囲】

【請求項1】 検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、複数の検索要求入力手段を備え、優先度の異なる複数の検索要求をユーザが入力できるようにしたことを特徴とする文書検索装置。

【請求項2】 検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書の複数のフィールドに対して、各々索引情報を持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段を備え、検索対象文書の順位付けに反映させる割合をフィールド毎にユーザが指定できるようにしたことを特徴とする文書検索装置。

【請求項3】 検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語数算定手段を備え、検索要求に含まれる複数の単語が検索対象文書中に同時に現れる場合に、当該文書に与える得点を加算することによりこれを優先的に表示させるようにしたことを特徴とする文書検索装置。

【請求項4】 検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語出現位置を索引に持ち、検索要求に含まれる複数の単語の検索対象文書中での出現位置の隣接度合を調べる単語近接度算定手段を備え、出現位置の近接度合によって、当該文書に与える得点を加算することによりこれを優先的に表示させるようにしたことを特徴とする文書検索装置。

【請求項5】 検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、複数の検索要求入力手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を備え、優先度の異なる複数の検索要求をユーザが入力できるようにすると共に単語共起関係が現れる文書に与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置。

【請求項6】 検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報をフィールド毎に索引に持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べるフィールド別単語共起関係照合手段を備え、検索対象文書の順位付けに反映させる割合をフィールド毎にユーザが指定できるようにすると共にフィールド毎に単語共起関係が現れる文書に

与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置。

【請求項7】 検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語算定手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を備え、検索要求に含まれる複数の単語が検索対象文書中に同時に現れる場合に、当該文書に与える得点を加算すると共に単語共起関係が現れる文書に与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は文書検索装置に関し、特に大量の文書データからユーザの入力した検索要求に合致する文書を探索し、その合致する度合によって順位付けを行なう文書検索装置に関するものである。

【0002】

【従来の技術】近年、文書検索の分野においては文書データベースの大規模化が進み、従来のようなキーワード検索や全文検索ではたとえ検索結果が高速に得られたとしても、その数が膨大で目的の文書を捜し出すのが困難な場合が増え、トータルな検索時間は必ずしも短縮されているとは言い難い。結果文書数を減らすためには、さらに別のキーワードを追加するなどして絞り込みを行なうという方法もあるが、目的とする文書が不必要な文書とともにふり落とされてしまわないような適切なキーワードを追加するのは難しい。

【0003】そこで、検索対象文書に文字列が存在するか否かだけでなく、その出現頻度等に注目して検索結果文書の順位付けを行ない、目的の文書を効率良く捜し出せるランキングの手法が注目されている。

【0004】図11は検索結果の順位付けを行なう従来の文書検索装置の構成を示したブロック図である。図11において、文書検索装置は、検索対象となる文書データ1101と、辞書1102と、辞書単語の文書中での出現頻度を格納した単語頻度索引1103と、文書データ1101から単語の出現頻度情報を得る単語頻度情報抽出手段1104と、ユーザからの検索要求を入力する検索要求入力手段1105と、単語頻度索引1103から単語の出現頻度を求める単語頻度算定手段1106と、単語の出現頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段1107と、頻度スコアをもとに各文書と検索要求の合致度合を示す文書スコアを算出する文書スコア算定手段1108と、文書スコアの順に文書を並べ替える文書順位付け手段1109と、スコアの順に並べられた結果文書を表示する検索結果表示手段1110とから構成されている。

【0005】図12は検索結果の順位付けを行なう従来の文書検索装置の索引作成の手順を示した流れ図である。まず、検索の前に予め単語頻度情報抽出手段1104が文書データ1101を調べ、総文書数、出現文書数とともに単語頻度索引1103に出力し単語頻度索引を作成しておく。

【0006】ステップ1201において、検索するユーザは検索要求入力手段1105により、検索要求を入力する。ステップ1202において、単語頻度算定手段1106*

$$SF_j = \sum_i (TF_{ij} \times IDF_i)$$

$$IDF_i = 1 - \log (ND_i / ND) \quad \dots (1)$$

ここで、 IDF_i は単語 W_i の全文書における偏りを表すパラメータである。

【0007】ステップ1204において、文書スコア算定手段1108が頻度スコア算定手段1107の出力した文書 D ※

$$S_j = SF_j$$

【0008】ステップ1205において、文書順位付け手段1109が文書スコア算定手段1108で算出された各文書のスコアの大きい順に検索結果を並べ替え、ステップ1206において、検索結果表示手段1110がその検索結果をユーザに提示する。

【0009】

【発明が解決しようとする課題】しかしながら前記の従来の構成では、図13に示すように、検索要求のうちの1つの単語が非常に高頻度で出現するような文書があった場合、ユーザの検索意図に反した文書であっても、高い順位にランクされてしまうという課題を有していた。また、検索対象文書の順位付けに用いるスコアの算出は、フィールドに関係なく文書単位で行なわれるため、その文書の内容を良く表しているであろうと思われる新聞記事の「見出し」や特許の「発明の名称」等の情報が活用できないという課題を有していた。

【0010】また、複数の検索要求を与える場合、それらに優先順位をつけられず、ユーザの要求を柔軟に表現できないという課題や、全てを含んで欲しい単語群を検索要求として与えた場合でも、1つの単語が非常に高頻度で出現する文書があると高順位にきてしまうという課題や、近接して出現しなければ意味の無いような単語群を検索要求として表現し、検索することができないといった課題を有していた。

【0011】本発明は前記従来技術の課題を解決するために、ユーザの検索要求を柔軟に受け付け、検索、順位付けを行なうことにより、ゴミの少ない高精度な検索結果が得られ、結果の絞り込みを含めた総合的な検索時間が短縮可能な文書検索装置を提供することを目的とする。

【0012】

【課題を解決するための手段】本発明の文書検索装置においては、検索対象文書中の単語共起関係の情報を索引

*は単語頻度索引1103を参照し、総文書数 ND および、検索要求入力手段1105で入力された検索要求に含まれる辞書単語 W_i ($i=1, 2, \dots, NW$: NW は検索要求に含まれる辞書単語数)について、当該単語の出現文書数 ND_i 、文書 D_j ($j=1, 2, \dots, ND$) 中での出現頻度 TF_{ij} を算出する。ステップ1203において、文書スコア算定手段1107が単語頻度算定手段1106からの出力をもとに文書 D_j の頻度スコア SF_j を式(1)によって算出する。

※ j の頻度スコア SF_j をもとに文書 D_j と検索要求の合致度合を示す文書スコア S_j を求める。従来の検索装置においては式(2)のように文書スコア S_j は頻度スコア SF_j そのものである。

$$\dots (2)$$

に持ち、検索要求に含まれる単語共起関係が検索対象文書に現れるか否かを調べる共起関係算定手段を備えたものである。

【0013】また、優先度の異なる複数の検索要求を入力するために、複数の検索要求入力手段を備えたものである。

【0014】また、検索対象文書の複数のフィールドに対して、各々索引情報を持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段を備えたものである。

【0015】また、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語数算定手段を備えたものである。

【0016】また、検索対象文書中の単語出現位置を索引に持ち、検索要求に含まれる複数の単語の検索対象文書中での出現位置が隣接もしくは比較的近いかどうかを調べる単語近接度算定手段を備えたものである。

【0017】本発明によれば、ユーザの検索要求を柔軟に表現し、検索、順位付けを行なうことにより、ユーザの検索意図に沿った、ゴミの少ない高精度な検索結果が得られ、結果の絞り込みを含めた総合的な検索時間が短縮可能な文書検索装置が得られる。

【0018】

【発明の実施の形態】本発明の請求項1記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、複数の検索要求入力手段を備え、優先度の異なる複数の検索要求をユーザが入力できるようにしたことを特徴とするものであり、ユーザの目的とする文書を高精度で検索することが可能となる作用を有する。

【0019】また、本発明の請求項2記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書の複数のフィ

ールドに対して、各々索引情報を持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段を備え、検索対象文書の順位付けに反映させる割合をフィールド毎にユーザが指定できるようにしたことを特徴とするものであり、ユーザの目的とする文書を高精度で検索することが可能となる作用を有する。

【0020】また、本発明の請求項3記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語算定手段を備え、検索要求に含まれる複数の単語が検索対象文書中に同時に現れる場合に、当該文書に与える得点を加算することによりこれを優先的に表示させるようにしたことを特徴とするものであり、ユーザの目的とする文書を高精度で検索することが可能となる作用を有する。

【0021】また、本発明の請求項4記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語出現位置を索引に持ち、検索要求に含まれる複数の単語の検索対象文書中での出現位置の近接度合を調べる単語近接度算定手段を備え、出現位置の近接度合によって、当該文書に与える得点を加算することによりこれを優先的に表示させるようにしたことを特徴とするものであり、ユーザの目的とする文書を高精度で検索することが可能となる作用を有する。

【0022】また、本発明の請求項5記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、複数の検索要求入力手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を備え、優先度の異なる複数の検索要求をユーザが入力できるようにすると共に単語共起関係が現れる文書に与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置としたものであり、ユーザの目的とする文書をより高精度で検索することが可能となる作用を有する。

【0023】また、本発明の請求項6記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報をフィールド毎に索引に持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べるフィールド別単語共起関係照合手段を備え、検索対象文書の順位付けに反映させる割合をフィールド毎にユーザが指定できるようにすると共にフィールド毎に単語共起関係が現れる文書に与える得点を加算

することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置としたものであり、ユーザの目的とする文書をより高精度で検索することが可能となる作用を有する。

【0024】また、本発明の請求項7記載の発明は、検索要求に基づいて、検索対象文書の検索と順位付けを行なう文書検索装置において、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語算定手段、および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を備え、検索要求に含まれる複数の単語が検索対象文書中に同時に現れる場合に、当該文書に与える得点を加算すると共に単語共起関係が現れる文書に与える得点を加算することにより、これを優先的に表示させるようにしたことを特徴とする文書検索装置としたものであり、ユーザの目的とする文書をより高精度で検索することが可能となる作用を有する。

【0025】以下、本発明の実施の形態について、図を参照しながら説明する。

【0026】(第1の実施の形態) 図1は本発明の第1の実施の形態における文書検索装置の構成を示したブロック図である。図1において、文書検索装置は、検索対象となる文書データ301と、辞書302と、辞書単語の検索対象文書中における出現頻度を格納した単語頻度索引303と、文書データ301から単語頻度情報を抽出し、単語頻度索引303に格納する単語頻度情報抽出手段304と、ユーザが優先させたいと思う検索要求を入力するための主題検索要求入力手段305と、主題検索要求に比べ優先度の低い検索要求をユーザが入力するための副題検索要求入力手段306と、主題検索要求入力手段305および副題検索要求入力手段306で入力された検索要求に含まれる辞書単語について、単語頻度索引303を調べ各文書中での出現頻度を求める単語頻度算定手段307と、単語頻度算定手段307で得られた単語頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段308と、頻度スコア算定手段308の出力をもとに各文書の文書スコアを算出する文書スコア算定手段309と、スコアの順に文書を並べ替える文書順位付け手段310と、スコアの順に並べられた結果文書を表示する検索結果表示手段311とから構成されている。

【0027】図2は本発明の第1の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【0028】まず、検索の前に予め、単語頻度情報抽出手段304が文書データ301を走査し、辞書302に含まれる辞書単語の各文書中での出現頻度を調べ、総文書数、当該単語の出現文書数とともに単語頻度索引303に出力し、索引を作成しておく。

【0029】ステップ401において、ユーザは主題検索要求入力手段305によって探したい文書に対する検索

10

20

30

40

50

要求でかつ重視してほしいものを主題検索要求として入力する。ステップ402において、ユーザは副題検索要求入力手段306によってそれほど重視しなくてもよいものを副題検索要求として入力する。

【0030】ステップ403において、単語頻度算定手段307は単語頻度索引303を参照し、総文書数NDを求めるとともに、主題検索要求入力手段305および副題検索要求入力手段306で入力された検索要求に含まれる辞書単語Wi (i=1、2、・・・、NW: NWは検索要求に含まれる辞書単語数) に対し、当該単語の文書Dj

(j=1、2、・・・、ND) 中での出現頻度TFij*

$SFj = \sum_i (TFij \times IDF_i \times WT_i)$

$IDF_i = 1 - \log (ND_i / ND)$

ここで、IDFiは単語Wiの全文書における偏りを表すパラメータである。

【0031】ステップ406において、文書スコア算定手段309は頻度スコア算定手段308の出力した頻度スコアから文書Djと検索要求の合致度合を示す文書スコアSjを前記式(2)によって算出する。

【0032】ステップ407において、文書順位付け手段310は文書スコア算定手段309の出力した各文書Djの文書スコアSjの値の大きい順に文書を並べ替え、ステップ407において、検索結果表示手段311が文書順位付け手段310の出力から得られたソート済の文書を検索の結果としてユーザに表示する。

【0033】前記のようにして、ユーザが入力する検索要求に優先度を設けることにより、ユーザの検索意図を柔軟に表現することができ、効率的な検索が可能となる。

【0034】(第2の実施の形態) 図3は本発明の第2の実施の形態における文書検索装置の構成を示したブロック図である。図3において、文書検索装置は、検索対象となる文書データ501と、辞書502と、辞書単語の検索対象文書中における出現頻度を各フィールド毎に格納したフィールド別単語頻度索引503と、文書データ501から単語頻度情報を抽出し、フィールド別単語頻度索引503に格納する単語頻度情報抽出手段504と、ユーザが検索要求を入力するための検索要求入力手段505と、文書のどのフィールドのスコアをどの程度文書スコアに反映させるかという割合をユーザが入力するためのフィールド割合入力手段506と、検索要求入力手段505で入力された検索要求に含まれる辞書単語について、フィールド別単語頻度索引503を調べ、各文書中でのフィールド毎の出現頻度を求めるフィールド別単語頻度算定手段507と、フィールド別単語頻度算定手段507で得られた単語頻度をもとに各文書のフィールド別の頻度スコアを求めるフィールド別頻度スコア算定手段508と、フィールド別頻度スコア算定手段508の出力とフィールド割合入力手段506からの入力をもとに各文書の文書スコアを算出する文

*と当該単語の出現する文書数NDiを取得し、頻度スコア算定手段308に出力する。ステップ404において、単語頻度算定手段307は単語Wiが主題検索要求、副題検索要求のどちらに含まれるかによってパラメータWTiを選び、ステップ405において、頻度スコア算定手段308は単語頻度算定手段307が出力した総文書数ND、単語Wiの出現する文書数NDi、文書Djにおける単語Wiの出現頻度TFij、単語Wiの重み付けパラメータWTiから、単語頻度による文書Djと検索要求の合致度合を示す頻度スコアSFjを式(3)によって算出する。

・・・(3)

書スコア算定手段509と、スコアの順に文書を並べ替える文書順位付け手段510と、スコアの順に並べられた結果文書を表示する検索結果表示手段511とから構成されている。

【0035】図4は本発明の第2の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【0036】まず、検索の前に予め、単語頻度情報抽出手段504が文書データ501を走査し、辞書502に含まれる辞書単語の各文書内でのフィールド毎の出現頻度を調べ、総文書数、当該単語の出現文書数とともに出現頻度をフィールド別単語頻度索引503に出力し、索引を作成しておく。

【0037】ステップ601において、ユーザは検索要求入力手段505によって探したい文書に対する検索要求を入力する。ステップ602において、ユーザはフィールド割合入力手段506によってフィールドFk (k=1、2、・・・、NF: NFは総フィールド数) のスコアを順位付けに反映させる割合Rkを入力する。

【0038】ステップ603において、フィールド別単語頻度算定手段507はフィールド別単語頻度索引503を参照し、総文書数NDを求めるとともに、検索要求入力手段505によって入力された検索要求に含まれる辞書単語Wi (i=1、2、・・・、NW: NWは検索要求に含まれる辞書単語数) に対し、当該単語の文書Dj (j=1、2、・・・、ND) のフィールドFk中での出現頻度TFijkとフィールドFkに当該単語の出現する文書数NDikを取得し、フィールド別頻度スコア算定手段508に出力する。

【0039】ステップ604において、フィールド別頻度スコア算定手段508はフィールド別単語頻度算定手段507が出力した総文書数ND、フィールドFkに単語Wiの出現する文書数NDik、文書DjのフィールドFkにおける単語Wiの出現頻度TFijk、単語頻度による文書DjのフィールドFkと検索要求の合致度合を示す頻度スコア頻度スコアSFjkを式(4)によって算出する。

$$SFjk = \sum_i (TFijk \times IDFik)$$

$$IDFik = 1 - \log (NDik / ND) \quad \dots (4)$$

【0040】ステップ605において、文書スコア算定手段509はフィールド別頻度スコア算定手段508の出力したフィールド毎の頻度とスコアフィールド割合入力手段*

$$Sj = \sum_k (SFjk \times Rk)$$

*506で入力されたフィールドFkを反映させる割合Rkから、文書Djと検索要求の合致度合を示す文書スコアSjを式(5)によって算出する。

$$\dots (5)$$

【0041】ステップ606において、文書順位付け手段510は文書スコア算定手段509の出力した各文書Djの文書スコアSjの値の大きい順に文書を並べ替え、ステップ607において、検索結果表示手段511が文書順位付け手段510の出力から得られたソート済の文書を検索の結果としてユーザに表示する。

【0042】前記のようにして、ユーザが検索対象フィールドのスコア配分の割合を変化させられるようにすることにより、ユーザの検索意図を柔軟に表現することができ、効率的な検索が可能となる。

【0043】(第3の実施の形態) 図5は本発明の第3の実施の形態における文書検索装置の構成を示したブロック図である。図5において、文書検索装置は、検索対象となる文書データ701と、辞書702と、辞書単語の検索対象文書中での出現頻度を格納した単語頻度索引703と、文書データ701から単語頻度情報を抽出し、単語頻度索引703に格納する単語頻度情報抽出手段705と、検索要求をユーザが入力するための検索要求入力手段707と、検索要求入力手段707で入力された検索要求に含まれる辞書単語について、単語頻度索引703を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段708と、単語頻度算定手段708で得られた単語頻度をもとに各文書のスコアを求める頻度スコア算定手段709と、単語頻度索引703を調べ、検索要求入力手段707で入力された検索要求に含まれる単語のうちいくつが、文書中に出現するかを求める出現語数算定手段710と、出現語数算定手段710で得られた出現語数に基づいて各文書に加算するスコアを求める出現語数スコア算定手段711と、頻度スコア算定手段709および出現語数スコア算定手段711の出力から各文書のスコアを算出する文書スコア算定手段712と、スコアの順に文書を並べ替える文書順位付け手段713と、スコアの順に並べられた結果文書を表示する検索結果表示手段714とから構成されている。

【0044】図6は本発明の第3の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【0045】まず、検索の前に予め、単語頻度情報抽出手段705が文書データ701を走査し、辞書702に含まれる辞書単語の各文書内での出現頻度を調べ、総文書数、当該単語の出現文書数とともに出現頻度を単語頻度索引70*

$$Sj = SFj + SAj \times \text{定数} \quad \dots (7)$$

【0052】出現語数スコアSAjを用意することにより、検索要求に含まれる単語をより多く含むような文書

※3に出力し、索引を作成しておく。

10 【0046】ステップ801において、ユーザは検索要求入力手段707によって探したい文書に対する検索要求を入力する。検索要求は複数の単語を入力してもよいし、文章を入力し別途単語抽出手段を用いて文章から単語を切り出すようにしてもよい。

【0047】ステップ802において、単語頻度算定手段708は単語頻度索引703を参照し、総文書数Nを求めるとともに検索要求入力手段707で入力された複数の辞書単語Wi (i=1, 2, ..., NW: NWは検索要求に含まれる辞書単語数) に対し、当該単語の文書Dj (j=1, 2, ..., ND) 中での出現頻度TFijと単語Wiの出現する文書数NDiを取得し、頻度スコア算定手段709に出力する。

【0048】ステップ803において、頻度スコア算定手段709は単語頻度算定手段708が出力した総文書数ND、単語Wiの出現する文書数NDi、文書Djにおける単語Wiの出現頻度TFijから、単語頻度による文書Djと検索要求の合致度合を示す頻度スコアSFjを前記式(1)によって算出する。

【0049】ステップ804において、出現語数算定手段710はステップ802までで既に得られている文書Djに出現する辞書単語の情報と検索要求入力手段707で入力された検索要求に含まれる複数の単語Wiを比較し、複数の単語Wiのうちで文書Djに出現するものの数NAjを算出し、出現語数スコア算定手段711に出力する。

【0050】ステップ805において、出現語数スコア算定手段711は出現語数算定手段710が出力した検索要求に含まれる単語のうちで文書Djに出現するものの数NAjに基づいた出現語数スコアSAjを算出する。例えば式(6)によって算出することができる。

$$SAj = NAj - 1 \quad \dots (6)$$

【0051】ステップ806において、文書スコア算定手段712は頻度スコア算定手段709が出力した頻度スコアSFjと出現語数スコア算定手段711が出力した出現語数スコアSAjから検索要求と文書Djの合致度合を表すスコアSjを式(7)によって算出する。

のスコアを高くし、優先的に表示させることが可能となる。また、式(7)において定数の値を変化させること

により、出現語数による優先表示の度合を変えることも可能である。

【0053】ステップ807において、文書順位付け手段713は文書スコア算定手段712が出力した各文書D_jの文書スコアS_jの値の大きい順に文書を並べ替える。ステップ808において、検索結果表示手段714は文書順位付け手段713の出力から得られたソート済の文書を検索の結果としてユーザに提示する。

【0054】前記のようにすれば、検索要求に複数の単語を含む場合に高頻度単語を1つだけ含むような文書が検索結果の上位に来てしまうというような不都合を回避でき、効率的な検索が可能となる。

【0055】(第4の実施の形態) 図7は本発明の第4の実施の形態における文書検索装置の構成を示したブロック図である。図7において、文書検索装置は、検索対象となる文書データ901と、辞書902と、辞書単語の検索対象文書中での出現頻度を格納した単語頻度索引903と、検索対象文書中に現れる単語の位置を格納した単語出現位置索引904と、文書データ901から単語頻度情報を抽出し、単語頻度索引903に格納する単語頻度情報抽出手段905と、文書データ901から単語の位置情報を求め、単語出現位置索引904に格納する単語出現位置情報抽出手段906と、検索要求をユーザが入力するための検索要求入力手段907と、検索要求入力手段907で入力された検索要求に含まれる辞書単語について、単語頻度索引903を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段908と、単語頻度算定手段908で得られた単語頻度をもとに各文書のスコアを求める頻度スコア算定手段909と、単語出現位置索引904を参照し、検索要求入力手段907で入力された検索要求に含まれる単語の文書中での出現位置を求める出現位置算定手段910と、単語出現位置算定手段910の出力から単語どうしの近接度合を求める単語近接度算定手段911と、単語近接度算定手段911の出力に基づいて各文書に加算するスコアを求める近接スコア算定手段912と、頻度スコア算定手段909および近接スコア算定手段912の出力から各文書のスコアを算出する文書スコア算定手段913と、スコアの順に文書を並べ替える文書順位付け手段914と、スコアの順に並べられた結果文書を表示する検索結果表示手段915とから構成されている。

【0056】図8は、本発明の第4の実施の形態における文書検索装置の検索の手順を示した流れ図である。ま*

$$NE_{jk} = 1 / (DST_{jk} + 1) \quad \dots (8)$$

【0063】なお、全ての組合せについて単語近接度を求めるのは計算コストがかかるため、閾値dを設け距離DST_{jk}がd以下であるような出現位置の組合せについてのみ計算をしたり、近接度を求める単語ペアをユーザが限定するようにしても良い。

* ず、検索の前に予め、単語頻度情報抽出手段905が文書データ901を走査し、辞書902に含まれる辞書単語の各文書内での出現頻度を調べ、総文書数、当該単語の出現文書数とともに出現頻度を単語頻度索引903に出力し、単語出現位置情報抽出手段906が辞書単語の各文書中での出現位置を調べ、単語出現位置索引904に出力し、索引を作成しておく。

【0057】ステップ1001において、ユーザは検索要求入力手段907によって探したい文書に対する検索要求として複数の単語を入力する。なお、検索要求としてユーザは文章を入力し、別途単語抽出手段を用いて文章から単語を切り出すようにしても良い。

【0058】ステップ1002において、単語頻度算定手段908は単語頻度索引903を参照し、総文書数Nを求めるとともに検索要求入力手段907で入力された複数の辞書単語W_i (i=1、2、・・・、NW: NWは検索要求に含まれる辞書単語数) に対し、文書D_j (j=1、2、・・・、ND) 中での出現頻度TF_{ij}と単語W_iの出現する文書数N_iを取得し、頻度スコア算定手段909に出力する。

【0059】ステップ1003において、単語出現位置算定手段910は単語出現位置索引904を参照し検索要求入力手段907で入力された複数の単語W_iの文書D_j中での出現位置を全て求め、単語近接度算定手段911に出力する。

【0060】ステップ1004において、頻度スコア算定手段909は単語頻度算定手段908が出力した総文書数ND、単語W_iの出現する文書数ND_i、文書D_jにおける単語W_iの出現頻度TF_{ij}から、単語頻度による文書D_jと検索要求の合致度合を示す頻度スコアSF_jを前記式(1)によって算出する。

【0061】ステップ1005において、単語近接度算定手段911は単語出現位置算定手段が出力した文書D_j中での各単語W_iの出現位置と単語長から、異なる単語の全ての出現位置の組合せP_k (k=1、2、・・・、NP: NPは異なる単語の全ての出現位置の組合せの数) について2単語の間の距離DST_{jk}を求め、ステップ1006において、DST_{jk}をもとに単語近接度NE_{jk}を求める。例えば単語近接度NE_{jk}は式(8)を用いて求めることができる。

【0062】

【0064】ステップ1007において、近接スコア算定手段912は単語近接度算定手段911の出力した単語近接度NE_{jk}により各文書D_jの近接スコアSN_jを式(9)により算出する。

$$S N j = \sum_k (N E j k)$$

【0065】ステップ1008において、文書スコア算定手段913は頻度スコア算定手段909が出力した頻度スコア $S F j$ と近接スコア算定手段912が出力した近接スコア $S j$ *

$$S j = S F j + S N j \times \text{定数}$$

【0066】このように、近接スコア $S N j$ を用意することにより、検索要求に含まれる異なり単語が互いに接近して出現するような文書のスコアを高くし、優先的に表示させることが可能となる。また、前記式(10)において定数の値を変化させることにより、単語近接度による優先表示の度合を変えることも可能である。

【0067】ステップ1009において、文書順位付け手段914は文書スコア算定手段913が出力した各文書 $D j$ の文書スコア $S j$ の値の大きい順に文書を並べ替える。ステップ1010において、検索結果表示手段915は文書順位付け手段914の出力から得られたソート済の文書を検索の結果としてユーザに提示する。

【0068】前記のようにすれば、検索要求に含まれる複数の単語が互いに近くに出現しなければ検索要求として意味をなさないような場合に、不要な文書が検索結果の上位に来てしまうというような不都合を回避でき、効率的な検索が可能となる。

【0069】(第5の実施の形態) 図9は本発明の第5の実施の形態における文書検索装置の構成を示したブロック図である。図9において、文書検索装置は、検索対象となる文書データ101と、辞書102と、辞書単語の検索対象文書中における出現頻度を格納した単語頻度索引103と、検索対象文書中に現れる単語共起情報を格納した単語共起索引104と、文書データ101から単語頻度情報を抽出し、単語頻度索引103に格納する単語頻度情報抽出手段105と、文書データ101から単語共起情報を抽出し、単語共起索引104に格納する単語共起情報抽出手段106と、ユーザが検索要求を入力するための検索要求入力手段107と、検索要求入力手段107で入力された検索要求に含まれる辞書単語について、単語頻度索引103を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段108と、単語頻度算定手段108で得られた単語頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段109と、検索要求入力手段107で入力された検索要求から単語共起情報を抽出する単語共起情報抽出手段110と、単語共起索引104の内容を参照し、単語共起情報抽出手段110が出力した検索要求に含まれる単語共起関係が、各文書にいくつ現れるかを求める単語共起関係照合手段111と、単語共起関係照合手段111によって得られた検索要求と文書に共通して出現する単語共起関係の度合によって各文書の共起スコアを求める共起スコア算定手段112と、頻度スコア算定手段109の出力と共起スコア算定手段112の出力から文書スコアを算出する文書スコア算定手段113と、スコアの順に文書を並べ替える文書順位付

... (9)

* $S N j$ から文書 $D j$ のスコア $S j$ 、すなわち検索要求と文書 $D j$ の合致度合を式(10)によって算出する。

... (10)

け手段114と、スコアの順に並べられた結果文書を表示する検索結果表示手段115とから構成されている。

【0070】図10は本発明の第5の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【0071】まず、検索の前に、予め単語頻度情報抽出手段105が文書データ101を走査し、総文書数、当該単語の出現文書数とともに単語頻度索引103に出力し、単語共起情報抽出手段106が文書データ101を走査し、各文書内での単語共起情報を求め、単語共起索引104に出力し、索引を作成しておく。単語共起情報としては例えば同一文章内に出現する単語のペアを共起関係にあると判断して抽出する方法や、形態素解析を行なって係受けの関係にある単語のペアを抽出する方法が考えられる。

【0072】ステップ201において、ユーザは検索要求入力手段107によって探したい文書に対する検索要求を文章で入力する。ステップ202において、単語頻度算定手段108は単語頻度索引103を参照し、総文書数 $N D$ を求めるとともに、検索要求入力手段107で入力された検索要求に含まれる辞書単語 $W i$ ($i = 1, 2, \dots$ 、 $N W$: $N W$ は検索要求に含まれる辞書単語数) に対し、当該単語の文書 $D j$ ($j = 1, 2, \dots$ 、 $N D$) 中での出現頻度 $T F i j$ と当該単語の出現する文書数 $N D i$ を取得し、頻度スコア算定手段109に出力する。

【0073】ステップ203において、頻度スコア算定手段109は単語頻度算定手段108が出力した総文書数 $N D$ 、単語 $W i$ の出現する文書数 $N D i$ 、文書 $D j$ における単語 $W i$ の出現頻度 $T F i j$ から、単語頻度による文書 $D j$ と検索要求による合致度合を示す頻度スコア $S F j$ を前記式(1)によって算出する。

【0074】ステップ204において、単語共起情報抽出手段110は検索要求入力手段107で入力された検索要求から、索引作成時と同様の方法によって単語共起関係 $C k$ ($k = 1, 2, \dots$ 、 $N C$: $N C$ は検索要求に含まれる単語共起関係の数) を抽出する。ステップ205において、単語共起関係照合手段111は単語共起索引104を参照し、文書 $D j$ に出現する単語共起関係のうち単語共起情報抽出手段110で得られた検索要求に含まれる単語共起関係 $C k$ と一致するものの数 $N C j$ を算出し、共起スコア算定手段112に出力する。

【0075】ステップ206において、共起スコア算定手段112は検索要求と文書の間で一致する単語共起関係の数に基づいて文書 $D j$ の共起スコア $S C j$ を算出する。最も単純な例としては式(11)のように共起の数をそのまま共起スコア $S C j$ とする。

$$SC_j = NC_j$$

【0076】ステップ207において、文書スコア算定手段113は頻度スコア算定手段109の出力した頻度スコアと共起スコア算定手段112の出力した共起スコアから文 *

$$S_j = SF_j + SC_j \times Const$$

【0077】ステップ208において、文書順位付け手段114は文書スコア算定手段113の出力した各文書 D_j の文書スコア S_j の値の大きい順に文書を並べ替え、ステップ209において、検索結果表示手段115が文書順位付け手段114の出力から得られたソート済の文書を検索

の結果としてユーザに表示する。
【0078】前記のようにして、単語頻度だけでなく検索要求と検索対象文書に含まれる単語共起関係を照合し、順位付けに反映させることにより、ユーザの検索意図により近い文書を検索結果の上位に表示することができ、効率的な検索が可能となる。

【0079】(第6の実施の形態) 図14は本発明の第6の実施の形態における文書検索装置の構成を示したブロック図である。図14において、文書検索装置は、検索対象となる文書データ1401と、辞書1402と、辞書単語の検索対象文書中における出現頻度を格納した単語頻度索引1403と、検索対象文書中に現れる単語共起情報を格納した単語共起索引1404と、文書データ1401から単語頻度情報を抽出し、単語頻度索引1403に格納する単語頻度情報抽出手段1405と、文書データ1401から単語共起情報を抽出し、単語共起索引1404に格納する単語共起情報抽出手段1406と、ユーザが重要視したいと思う検索要求を入力するための主題検索要求入力手段1407と、ユーザが主題検索要求に比べそれほど重要視しなくても良いと思う検索要求を入力するための副題検索要求入力手段1408と、主題検索要求入力手段1407および副題検索要求入力手段1408で入力された検索要求に含まれる辞書単語について、単語頻度索引1403を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段1409と、単語頻度算定手段1409で得られた単語頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段1410と、主題検索要求入力手段1407および副題検索要求入力手段1408で入力された検索要求から単語共起情報を抽出する単語共起情報抽出手段1411と、単語共起索引1404の内容を参照し、単語共起情報抽出手段1411が出力した検索要求に含まれる単語共起関係が、各文書にいくつ現れるかを求める単語共起関係照合手段1412と、単語共起関係照合手段1412によって得られた検索要求と文書に共通して出現する単語共起関係の数によって各文書の共起スコアを求める共起スコア算定手段1413と、頻度スコア算定手段1410の出力と共起スコア算定手段1413の出力から各文書に対する最終的なスコアを算出する文書スコア算定手段1414と、スコアの順に文書を並べ替える文書順位付け手段1415と、スコアの順に並べられた結果文書を表示する検索結果表示手段1416とから構成される。

... (11)

* 書 D_j と検索要求の合致度合を示す文書スコア S_j を式(12)によって算出する。

... (12)

【0080】図15、図16、図17および図18は本発明の第6の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【0081】まず、検索の前に予め、単語頻度情報抽出手段1405が文書データ1401を走査し、辞書1402に含まれる辞書単語の各文書内での出現頻度を調べ、総文書数、当該単語の出現文書数とともに単語頻度索引1403に出力し、単語共起情報抽出手段1406が文書データ1401を走査し、各文書内での単語共起情報を求め、単語共起索引1404に出力し、索引を作成しておく。単語共起情報としては例えば同一文章内に出現する単語のペアを共起関係にあると判断して抽出する方法や、形態素解析を行なって係受けの関係にある単語のペアを抽出する方法が考えられる。

【0082】ステップ1501において、ユーザは主題検索要求入力手段1407によって探したい文書に対する検索要求でかつ重視したいものを主題検索要求として入力する。

【0083】ステップ1502において、ユーザは副題検索要求入力手段1408によって主題検索要求に比べそれほど重視しなくてもよいものを副題検索要求として入力する。

【0084】ステップ1503において、単語頻度算定手段1409は単語頻度索引1403を参照し、総文書数 ND を求めるとともに、主題検索要求入力手段1407および副題検索要求入力手段1408で入力された検索要求に含まれる辞書単語 W_i ($i=1, 2, \dots, NW$: NW は検索要求に含まれる辞書単語数) に対し、当該単語の文書 D_j ($j=1, 2, \dots, ND$) 中での出現頻度 TF_{ij} と当該単語の出現する文書数 ND_i を取得し、ステップ1504において、単語頻度算定手段1409は単語 W_i が主題検索要求、副題検索要求のどちらに含まれるかによって重み付けパラメータ WT_i を選び、頻度スコア算定手段1410に出力する。

【0085】ステップ1505において、頻度スコア算定手段1410は単語頻度算定手段1409が出力した総文書数 ND 、単語 W_i の出現する文書数 ND_i 、文書 D_j における単語 W_i の出現頻度 TF_{ij} 、単語 W_i の重み付けパラメータ WT_i から、単語頻度による文書 D_j と検索要求の合致度合いを示す頻度スコア SF_j を前記式

(5)によって算出し、文書スコア算定手段1414に出力する。

【0086】ステップ1506において、単語共起情報抽出手段1411は索引作成時と同様の方法によって主題検索要求入力手段1407で入力された主題検索要求から主題

共起関係 Csk ($k=1, 2, \dots, NCs:NCs$ は主題検索要求に含まれる単語共起関係の数) を抽出し、単語共起関係照合手段1412に出力する。

【0087】ステップ1507において、単語共起関係照合手段1412は単語共起索引1404を参照し、文書 Dj に出現する単語共起関係のうち単語共起情報抽出手段1411で得られた主題共起関係 Csk と一致するものの数 $NCsj$ を算出し、共起スコア算定手段1413に出力する。

【0088】ステップ1508において、単語共起情報抽出手段1411は索引作成時と同様の方法によって副題検索要求入力手段1408で入力された副題検索要求から副題共起関係 Cfm ($m=1, 2, \dots, Ncf:Ncf*$

$$SCj = NCsj \times (Ncf + 1) + Ncfj \quad \dots (13)$$

【0091】ステップ1511において、文書スコア算定手段1414は式(14)に基づいて頻度スコアの最大値※

$$SR = \text{Max}(SFj) - \text{Min}(SFj) \quad \dots (14)$$

【0092】ステップ1512において、文書スコア算定手段1414は頻度スコア算定手段1410の出力した頻度スコアと共起スコア算定手段1413の出力した共起スコア★

$$Sj = SFj + SCj \times SR$$

【0093】ステップ1513において、文書順位付け手段1415は文書スコア算定手段1414の出力した各文書 Dj の文書スコア Sj の値の大きい順に文書を並べ替え、ステップ1514において検索結果表示手段1416が文書順位付け手段1415の出力から得られたソート済みの文書を検索の結果としてユーザに表示する。前記のようにして、主題検索要求と副題検索要求という重要視する度合の異なる検索要求を受け付け、検索要求と文書の合致度合いを判定する基準として、主題共起関係>副題共起関係>主題単語頻度>副題単語頻度、の順に優先することにより、ユーザの検索意図により近い文書を検索結果の上位に表示することができ、高精度で効率的な検索が可能となる。

【0094】(第7の実施の形態) 図19は本発明の第7の実施の形態における文書検索装置の構成を示したブロック図である。図19において、文書検索装置は、検索対象となる文書データ1901と、辞書1902と、辞書単語の検索対象文書中における出現頻度をフィールド毎に格納したフィールド別単語頻度索引1903と、検索対象文書中に現れる単語共起情報をフィールド毎に格納したフィールド別単語共起索引1904と、文書データ1901から単語頻度情報を抽出し、フィールド別単語頻度索引1903に格納する単語頻度情報抽出手段1905と、文書データ1901から単語共起情報を抽出し、フィールド別単語共起索引1904に格納する単語共起情報抽出手段1906と、ユーザが検索要求を入力するための検索要求入力手段1907と、検索要求入力手段1907で入力された検索要求に含まれる辞書単語について、フィールド別単語頻度索引1903を調べ当該単語の文書中でのフィールド毎の出現頻度を求めるフィールド別単語頻度算定手段1908と、フィールド別単語

*は副題検索要求に含まれる単語共起関係の数) を抽出し、単語共起関係照合手段1412に出力する。

【0089】ステップ1509において、単語共起関係照合手段1412は単語共起索引1404を参照し、文書 Dj に出現する単語共起関係のうち単語共起情報抽出手段1411で得られた副題共起関係 Cfm と一致するものの数 $Ncfj$ を算出し、共起スコア算定手段1413に出力する。

【0090】ステップ1510において、共起スコア算定手段1413は式(13)に基づいて文書 Dj の共起スコア SCj を算出し、文書スコア算定手段1414に出力する。

※と最小値の差 SR を算出する。

★ら文書 Dj と検索要求との合致度合いを示す文書スコア Sj を式(15)によって算出する。

$$\dots (15)$$

頻度算定手段1908で得られた単語頻度をもとに各文書のフィールド毎の頻度スコアを求めるフィールド別頻度スコア算定手段1909と、検索要求入力手段1907で入力された検索要求から単語共起情報を抽出する単語共起情報抽出手段1910と、フィールド別単語共起索引1904の内容を参照し、単語共起情報抽出手段1910が出力した検索要求に含まれる単語共起関係が、各文書の各フィールドにいくつ現れるかを求めるフィールド別単語共起関係照合手段1911と、フィールド別単語共起関係照合手段1911によって得られた検索要求と文書の各フィールドに共通して出現する単語共起関係の数によって各文書のフィールド毎の共起スコアを求めるフィールド別共起スコア算定手段1912と、各フィールドのスコアをどの程度文書の順位付けにスコアに反映させるかという割合をユーザが入力するためのフィールド割合入力手段1913と、フィールド別頻度スコア算定手段1909の出力とフィールド別共起スコア算定手段1912の出力とフィールド割合入力手段1913の出力から各文書に対する最終的なスコアを算出する文書スコア算定手段1914と、スコアの順に文書を並べ替える文書順位付け手段1915と、スコアの順に並べられた結果文書を表示する検索結果表示手段1916とから構成される。

【0095】図20、図21、図22および図23は本発明の第7の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【0096】まず、検索の前に予め、単語頻度情報抽出手段1905が文書データ1901を走査し、辞書1902に含まれる辞書単語の各文書内でのフィールド毎の出現頻度を調べ、総文書数、当該単語の出現文書数とともにフィールド別単語頻度索引1903出力し、単語共起情報抽出手段19

06が文書データ1901を走査し、各文書内でのフィールド毎の単語共起情報を求め、フィールド別単語共起索引1904に出力し、索引を作成しておく。単語共起情報としては例えば同一文章内に出現する単語のペアを共起関係にあると判断して抽出する方法や、形態素解析を行なって係受けの関係にある単語のペアを抽出する方法が考えられる。

【0097】ステップ2001において、ユーザは検索要求入力手段1907によって探したい文書に対する検索要求を入力する。

【0098】ステップ2002において、ユーザはフィールド割合入力手段1913によってフィールドF_m (m=1、2、・・・、NF：NFは総フィールド数) のスコアを順位付けに反映させる割合R_mを入力する。

【0099】ステップ2003において、フィールド別単語頻度算定手段1908はフィールド別単語頻度索引1903を参照し、総文書数NDを求めるとともに、検索要求入力手段1907で入力された検索要求に含まれる辞書単語W_i (i=1、2、・・・、NW：NWは検索要求に含まれる辞書単語数) に対し、当該単語の文書D_j (j=1、2、・・・、ND) のフィールドF_m中での出現頻度TF_{ijm}とフィールドF_mに当該単語の出現する文書数ND_{im}を取得し、フィールド別頻度スコア算定手段1909に出力する。

$$SC_{jm} = NC_{jm}$$

【0104】ステップ2008において、文書スコア算定手段1914は式(17)に基づいてフィールド別頻度スコア

$$SR = \text{Max}(SF_{jm}) - \text{Min}(SF_{jm}) \quad \dots (17)$$

【0105】ステップ2009において、文書スコア算定手段1914はフィールド別頻度スコア算定手段1909の出力したフィールド毎の頻度スコアSF_{jm}とフィールド別共起スコア算定手段1912の出力したフィールド毎の共

$$S_j = \sum_m ((SF_{jm} + SC_{jm} \times SR) \times R_m) \quad \dots (18)$$

【0106】ステップ2010において文書順位付け手段1915は文書スコア算定手段1914の出力した各文書D_jの文書スコアS_jの値の大きい順に文書を並べ替え、ステップ2011において検索結果表示手段1916が文書順位付け手段1915の出力から得られたソート済みの文書を検索の結果としてユーザに表示する。

【0107】前記のようにして、ユーザが検索対象フィールドのスコア配分の割合を変化させられるようにすることにより、ユーザの検索意図を柔軟に表現することができ、効率的な検索が可能となる。

【0108】(第8の実施の形態) 図24は本発明の第8の実施の形態における文書検索装置の構成を示したブロック図である。図24において、文書検索装置は、検索対象となる文書データ2401と、辞書2402と、辞書単語の検索対象文書中における出現頻度を格納した単語頻度索引2403と、検索対象文書中に現れる単語共起情報を格

【0100】ステップ2004において、フィールド別頻度スコア算定手段1909はフィールド別単語頻度算定手段1908が出力した総文書数ND、フィールドF_mに単語W_iの出現する文書数ND_{im}、文書D_jのフィールドF_mにおける単語W_iの出現頻度TF_{ijm}から、単語頻度に基づく文書D_jのフィールドF_mと検索要求の合致度合いを示す頻度スコアSF_{jm}を前記式(4)によって算出し、文書スコア算定手段1914に出力する。

【0101】ステップ2005において、単語共起情報抽出手段1910は索引作成時と同様の方法によって検索要求入力手段1907で入力された検索要求から共起関係C_k (k=1、2、・・・、NC：NCは検索要求に含まれる単語共起関係の数) を抽出し、フィールド別単語共起関係照合手段1911に出力する。

【0102】ステップ2006において、フィールド別単語共起関係照合手段1911はフィールド別単語共起索引1904を参照し、文書D_jのフィールドF_mに出現する単語共起関係のうち単語共起情報抽出手段1910で得られた単語共起関係C_kと一致するものの数NC_{jm}を算出し、フィールド別共起スコア算定手段1912に出力する。

【0103】ステップ2007において、フィールド別共起スコア算定手段1912は式(16)に基づいて文書D_jのフィールドF_mの共起スコアSC_{jm}を算出し、文書スコア算定手段1914に出力する。

$$\dots (16)$$

※コアの最大値と最小値の差SRを算出する。

★起スコアSC_{jm}とフィールド割合入力手段で入力されたスコア配分割合R_mから文書D_jと検索要求との合致度合いを示す文書スコアS_jを式(18)によって算出する。

納した単語共起索引2404と、文書データ2401から単語頻度情報を抽出し、単語頻度索引2403に格納する単語頻度情報抽出手段2405と、文書データ2401から単語共起情報を抽出し、単語共起索引2404に格納する単語共起情報抽出手段2406と、ユーザが検索要求を入力するための検索要求入力手段2407と、検索要求入力手段2407で入力された検索要求に含まれる辞書単語について、単語頻度索引2403を調べ当該単語の文書中での出現頻度を求める単語頻度算定手段2408と、単語頻度算定手段2408で得られた単語頻度をもとに各文書の頻度スコアを求める頻度スコア算定手段2409と、単語頻度索引2403を調べ、検索要求入力手段2407で入力された検索要求に含まれる辞書単語が、各文書中にいくつ出現するのかを求める出現語数算定手段2410と、出現語数算定手段2411で得られた出現語数をもとに各文書の出現語数スコアを求める出現語数スコア算定手段2411と、検索要求入力手段2407で入力され

た検索要求から単語共起情報を抽出する単語共起情報抽出手段2412と、単語共起索引2404の内容を参照し、単語共起情報抽出手段2412が出力した検索要求に含まれる単語共起関係が、各文書にいくつ現れるかを求める単語共起関係照合手段2413と、単語共起関係照合手段2413によって得られた検索要求と文書に共通して出現する単語共起関係の数によって各文書の共起スコアを求める共起スコア算定手段2414と、頻度スコア算定手段2409の出力と出現語数スコア算定手段2411の出力と共起スコア算定手段2414の出力から各文書に対する最終的なスコアを算出する文書スコア算定手段2415と、スコアの順に文書を並べ替える文書順位付け手段2416と、スコアの順に並べられた結果文書を表示する検索結果表示手段2417とから構成される。

【0109】図25、図26、図27、図28および図29は本発明の第8の実施の形態における文書検索装置の検索の手順を示した流れ図である。

【0110】まず、検索の前に予め、単語頻度情報抽出手段2405が文書データ2401を走査し、辞書2402に含まれる辞書単語の各文書内での出現頻度を調べ、総文書数、当該単語の出現文書数とともに単語頻度索引2403に出力し、単語共起情報抽出手段2406が文書データ2401を走査し、各文書内での単語共起情報を求め、単語共起索引2404に出力し、索引を作成しておく。単語共起情報としては例えば同一文章内に出現する単語のペアを共起関係にあると判断して抽出する方法や、形態素解析を行なって係受けの関係にある単語のペアを抽出する方法が考えられる。

【0111】ステップ2501において、ユーザは検索要求入力手段2407によって探したい文書に対する検索要求を入力する。

【0112】ステップ2502において、単語頻度算定手段2408は単語頻度索引2403を参照し、総文書数NDを求めるとともに、検索要求入力手段2407で入力された検索要求に含まれる辞書単語Wi (i=1、2、・・・、NW: NWは検索要求に含まれる辞書単語数) に対し、当該単語の文書Dj (j=1、2、・・・、ND) 中での出現頻度TFijと当該単語の出現する文書数NDiを取得し、頻度スコア算定手段2409に出力する。

【0113】ステップ2503において、頻度スコア算定手段2409は単語頻度算定手段2408が出力した総文書数*

$$S_j = S F_j + (S A_j + S C_j \times N W) \times S R \quad \cdots (19)$$

【0121】ステップ2511において、文書順位付け手段2416は文書スコア算定手段2415の出力した各文書Djの文書スコアSjの値の大きい順に文書を並べ替え、ステップ2512において検索結果表示手段2417が文書順位付け手段2416の出力から得られたソート済みの文書を検索の結果としてユーザに表示する。

【0122】前記のようにして、検索要求と文書の合致度合いを判定する基準として、単語頻度だけでなく、共

*ND、単語Wiの出現する文書数NDi、文書Djにおける単語Wiの出現頻度TFijから、単語頻度による文書Djと検索要求の合致度合いを示す頻度スコアSFjを前記式(1)によって算出し、文書スコア算定手段2415に出力する。

【0114】ステップ2504において、出現語数算定手段2410は単語頻度索引2403を参照し、検索要求入力手段2407で入力された検索要求に含まれる辞書単語Wiのうち、文書Djに出現する単語の数NAjを算出し、出現語数スコア算定手段2411に出力する。

【0115】ステップ2505において、出現語数スコア算定手段2411は出現語数算定手段2410の出力した出現語数NAjにもとづいて、文書Djの出現語数スコアを前記式(6)によって算出し、出現語数スコア算定手段2411に出力する。

【0116】ステップ2506において、単語共起情報抽出手段2412は索引作成時と同様の方法によって検索要求入力手段2407で入力された検索要求から共起関係Ck (k=1、2、・・・、NC: NCは検索要求に含まれる単語共起関係の数) を抽出し、単語共起関係照合手段2413に出力する。

【0117】ステップ2507において、単語共起関係照合手段2413は単語共起索引2404を参照し、単語共起情報抽出手段2404で得られた各単語共起関係Ckが出現する文書を求め、単語共起関係Ckのうちで文書Djに出現するものの数NCjを算出し、共起スコア算定手段2414に出力する。

【0118】ステップ2508において、共起スコア算定手段2414は前記式(11)に基づいて文書Djの共起スコアSCjを算出し、文書スコア算定手段2415に出力する。

【0119】ステップ2509において、文書スコア算定手段2415は前記式(14)に基づいて頻度スコアの最大値と最小値の差SRを算出する。

【0120】ステップ2510において、文書スコア算定手段2415は頻度スコア算定手段2409の出力した頻度スコアSFjと出現語数スコア算定手段2411の出力した出現語数スコアSAjと共起スコア算定手段2414の出力した共起スコアSCjから文書Djと検索要求との合致度合いを示す文書スコアSjを式(19)によって算出する。

起関係、出現語数を採り入れ、共起関係>出現語数>単語頻度、の順に優先することにより、ユーザの検索意図により近い文書を検索結果の上位に表示することができ、高精度で効率的な検索が可能となる。

【0123】

【発明の効果】以上のように本発明の文書検索装置においては、優先度の異なる複数の検索要求を入力するため複数の検索要求入力手段を設けることにより、また、

10

20

30

40

50

検索対象文書のフィールド毎の索引情報を持ち、順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段を設けることにより、また、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語数算定手段を設けることにより、また、検索対象文書中の単語出現位置を索引に持ち、検索要求に含まれる複数の単語の検索対象文書中の出現位置が隣接もしくは比較的近いかどうかを調べる単語近接度算定手段を設けることにより、また、検索対象文書中の単語共起情報を索引に持ち、複数の検索要求入力手段および検索要求に含まれる単語共起関係が検索対象文書に現れるか否かを調べる単語共起関係照合手段を設けることにより、また、検索対象文書中の単語出現頻度と単語共起情報をフィールド毎に索引に持ち、検索対象文書の順位付けに反映させるフィールドの割合をユーザが指定できるフィールド割合入力手段および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べるフィールド別単語共起関係照合手段を設けることにより、また、検索対象文書中の単語出現頻度と単語共起情報を索引に持ち、検索要求に含まれる複数の単語が検索対象文書中にいくつ含まれるかを算出する出現語数算定手段および検索要求に含まれる単語共起関係が検索対象文書中に現れるかどうかを調べる単語共起関係照合手段を設けることにより、ユーザの検索要求を柔軟に受け付け、検索、順位付けを行なうことにより、ゴミの少ない高精度な検索結果が得られ、結果の絞り込みを含めた総合的な検索時間が短縮可能な文書検索装置が得られるものである。

【図面の簡単な説明】

【図1】本発明の第1の実施の形態における文書検索装置の構成を示すブロック図、

【図2】本発明の第1の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図3】本発明の第2の実施の形態における文書検索装置の構成を示すブロック図、

【図4】本発明の第2の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図5】本発明の第3の実施の形態における文書検索装置の構成を示すブロック図、

【図6】本発明の第3の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図7】本発明の第4の実施の形態における文書検索装置の構成を示すブロック図、

【図8】本発明の第4の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図9】本発明の第5の実施の形態における文書検索装置の構成を示すブロック図、

【図10】本発明の第5の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図11】従来の文書検索装置の構成を示すブロック

図、

【図12】従来の文書検索装置の検索の手順を示す流れ図、

【図13】従来の文書検索装置の検索の例を示す図、

【図14】本発明の第6の実施の形態における文書検索装置の構成を示すブロック図、

【図15】本発明の第6の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図16】本発明の第6の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図17】本発明の第6の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図18】本発明の第6の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図19】本発明の第7の実施の形態における文書検索装置の構成を示すブロック図、

【図20】本発明の第7の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図21】本発明の第7の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図22】本発明の第7の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図23】本発明の第7の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図24】本発明の第8の実施の形態における文書検索装置の構成を示すブロック図、

【図25】本発明の第8の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図26】本発明の第8の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図27】本発明の第8の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図28】本発明の第8の実施の形態における文書検索装置の検索の手順を示す流れ図、

【図29】本発明の第8の実施の形態における文書検索装置の検索の手順を示す流れ図である。

【符号の説明】

101、301、501、701、901、1101、1401、1901、2401 文書データ

102、302、502、702、902、1102、1402、1902、2402 辞書

103、303、503、703、903、1103、1403、2403 単語頻度索引

104、1404、2404 単語共起索引

105、304、504、705、905、1104、1405 単語頻度情報抽出手段

1905、2405 単語頻度情報抽出手段

106、1406、1906、2406 単語共起情報抽出手段

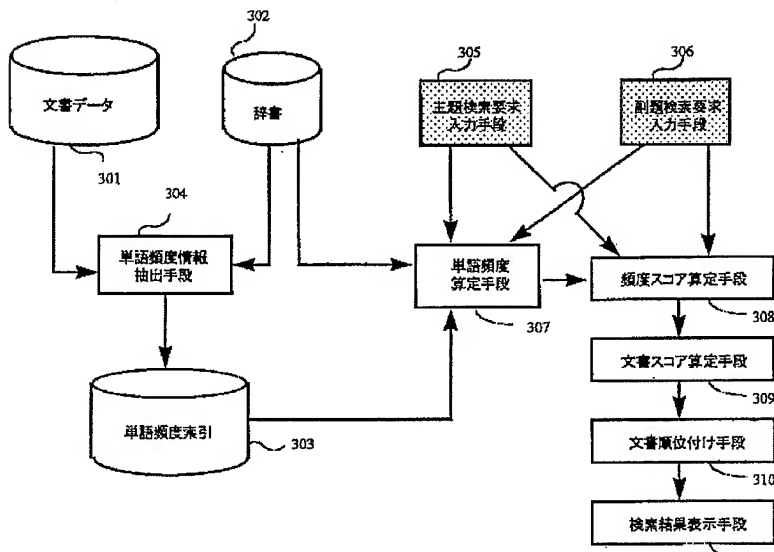
107、505、707、907、1105、1907、2407 検索要求入力手段

108、307、708、908、1106、1409、2408 単語頻度算定手段
 109、308、709、909、1107、1410、2409 頻度スコア算定手段
 110、1406、1411、1906、1910、2406、2412 単語共起情報抽出手段
 111、1412、2413 単語共起関係照合手段
 112、1413、2414 共起スコア算定手段
 113、309、509、712、913、1108、1414 文書スコア算定手段
 1914、2415 文書スコア算定手段
 114、310、510、713、914、1109、1415 文書順位付け手段
 1915、2416 文書順位付け手段
 115、311、511、714、915、1110、1416 検索結果表示手段

* 1916、2417 検索結果表示手段
 305、1407 主題検索要求入力手段
 306、1408 副題検索要求入力手段
 506、1913 フィールド割合入力手段
 507、1908 フィールド別単語頻度算定手段
 508、1909 フィールド別頻度スコア算定手段
 710、2410 出現語数算定手段
 711、2411 出現語数スコア算定手段
 904 単語出現位置索引
 10 906 単語出現位置情報抽出手段
 910 単語出現位置算定手段
 911 単語近接度算定手段
 912 近接スコア算定手段
 1911 フィールド別単語共起関係照合手段
 1912 フィールド別共起スコア算定手段

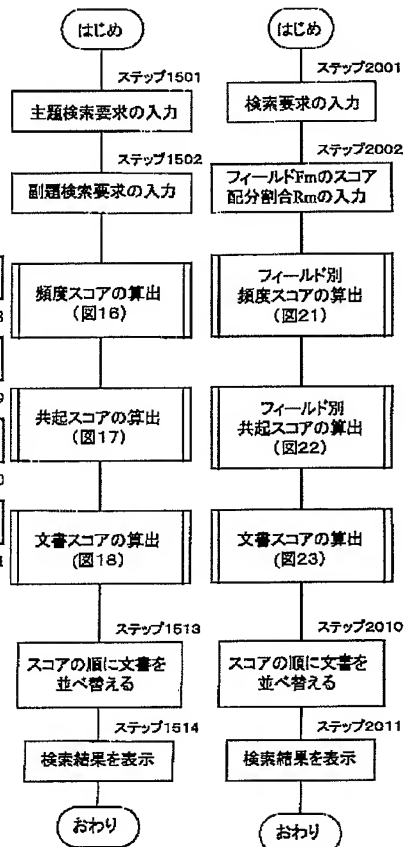
*

【図1】

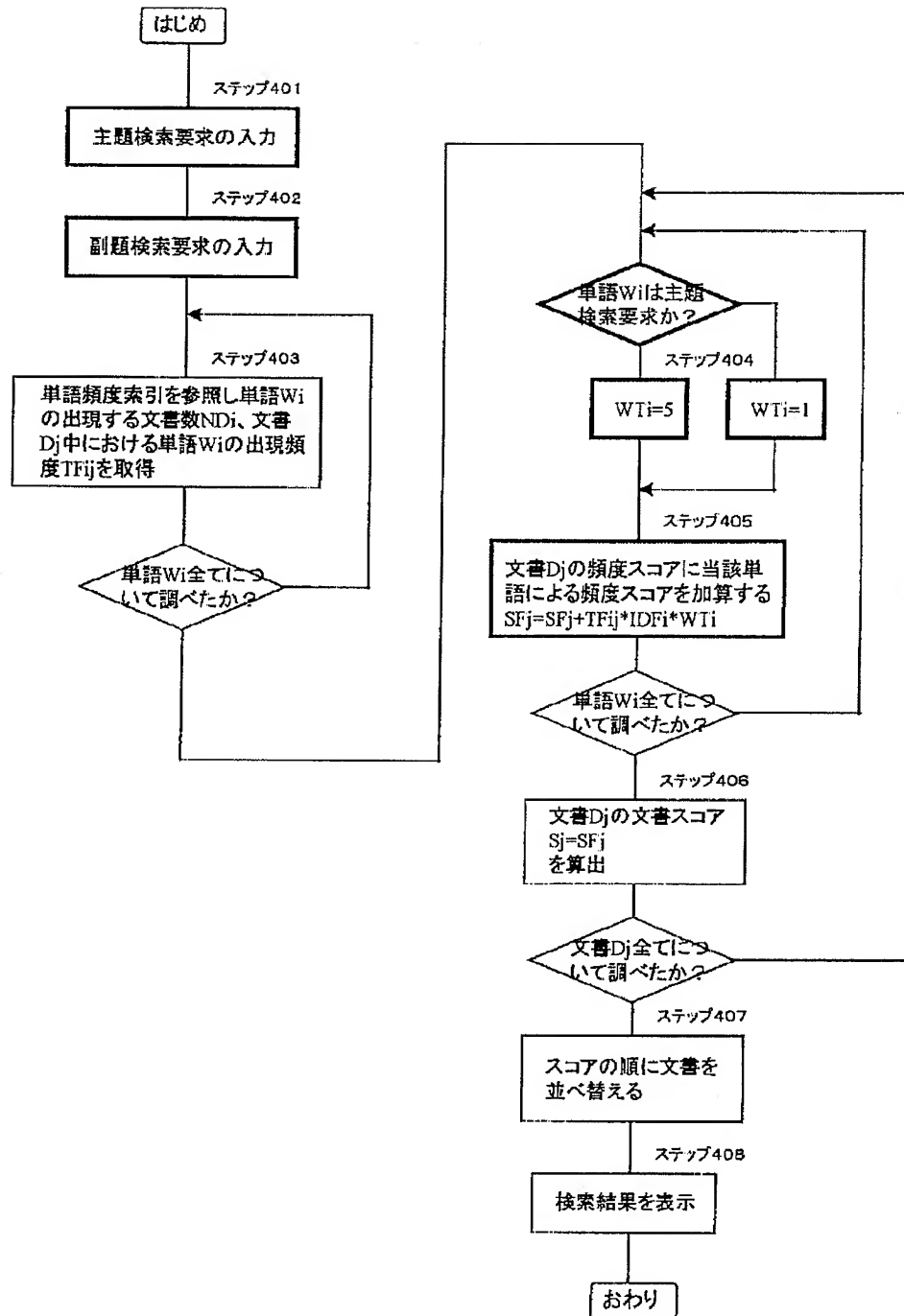


【図15】

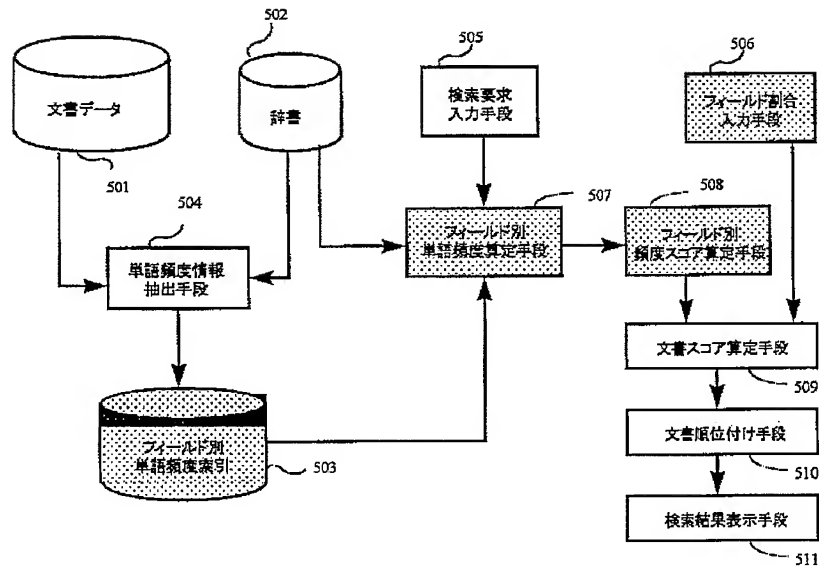
【図20】



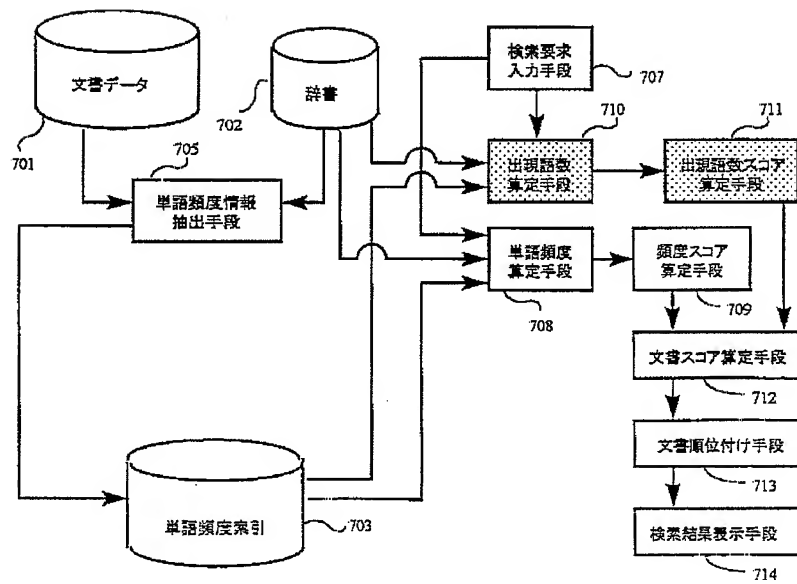
【図2】



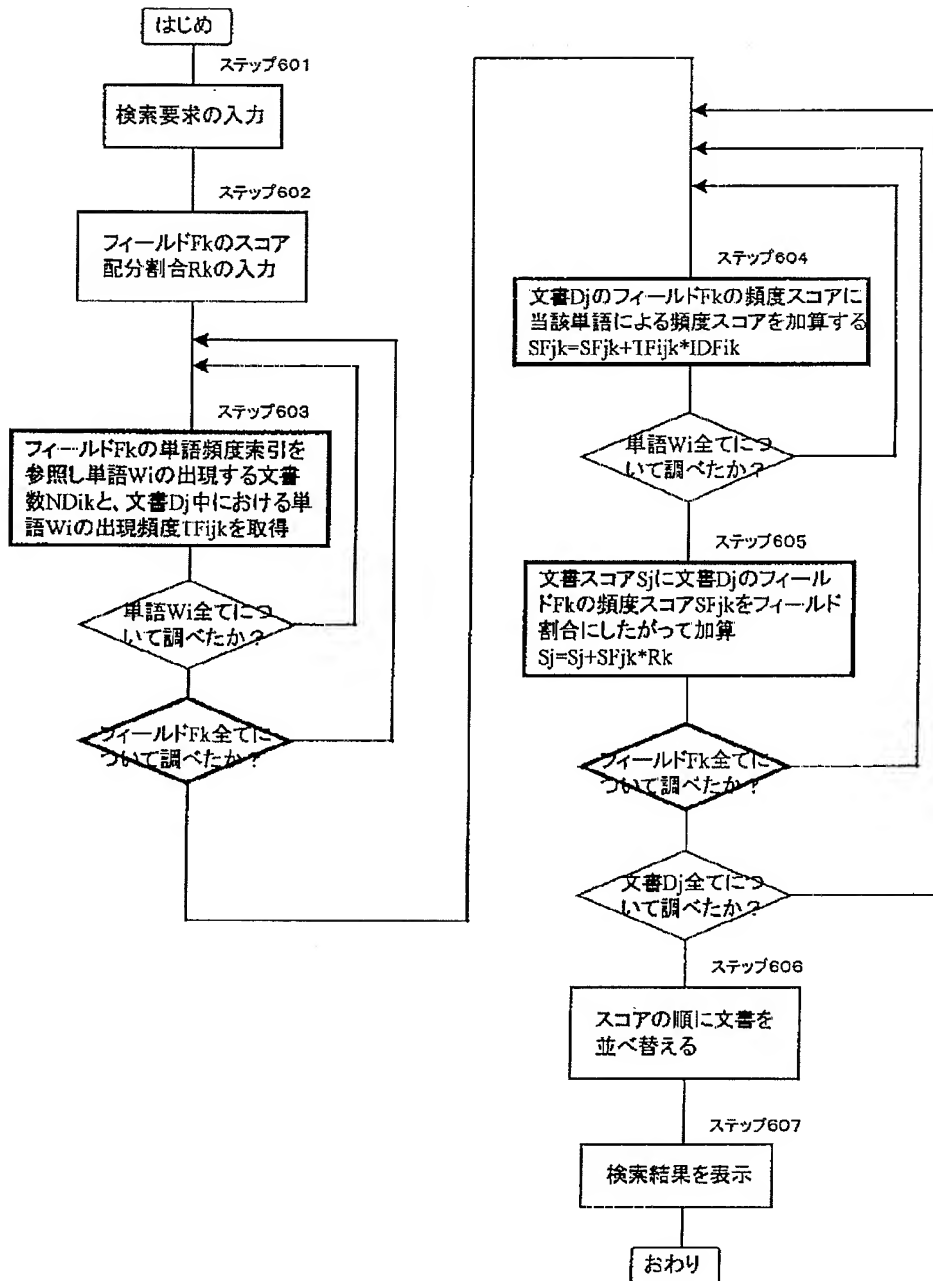
【図3】



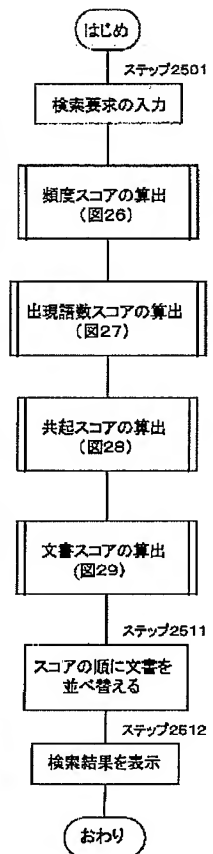
【図5】



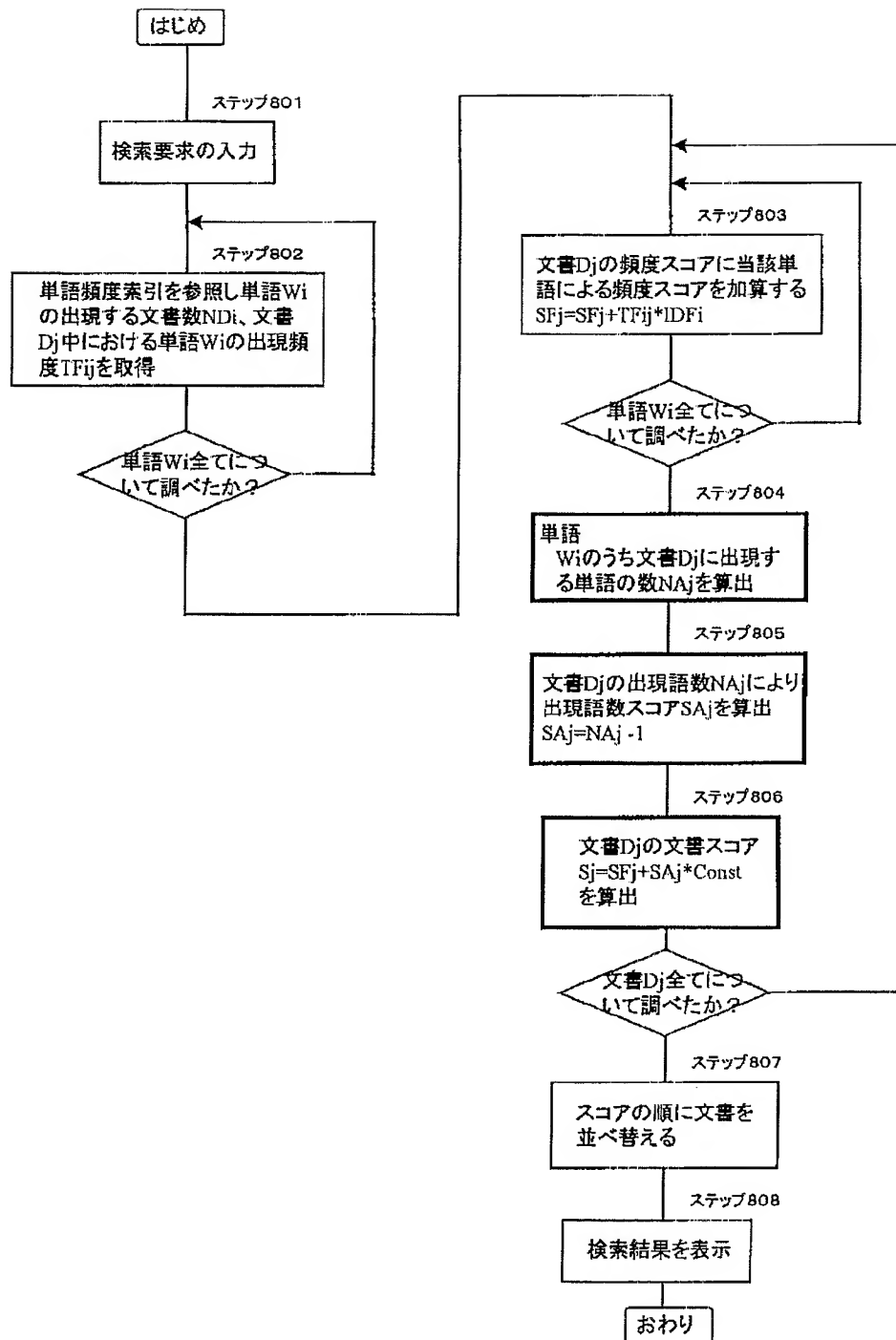
【図4】



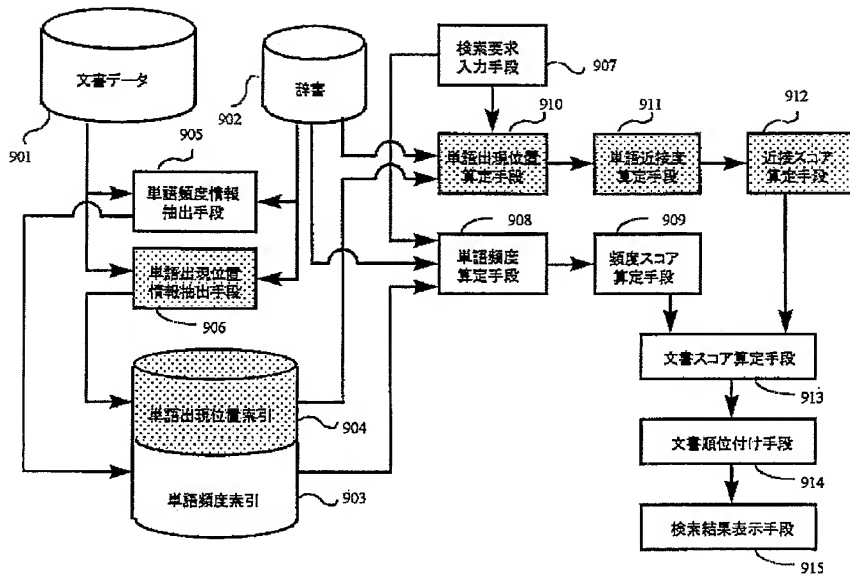
【図25】



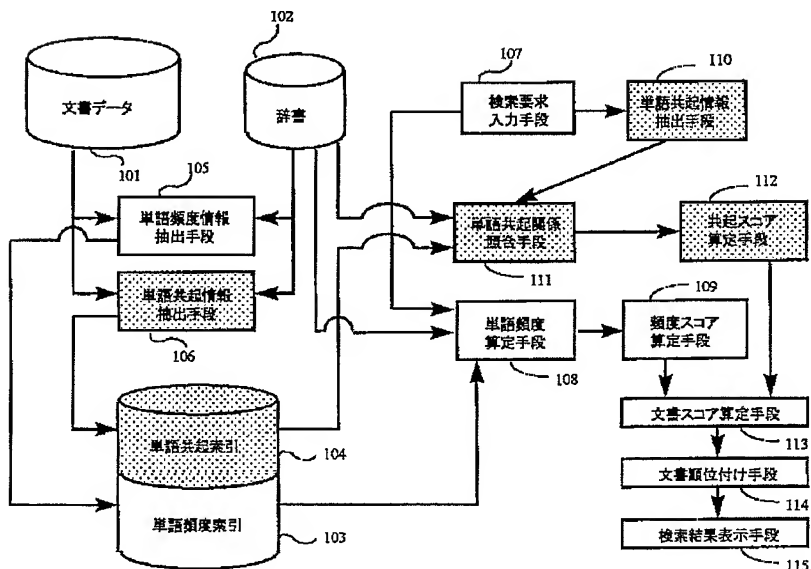
【図6】



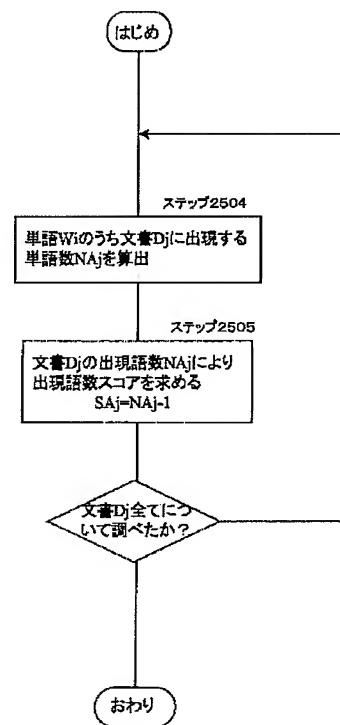
【図7】



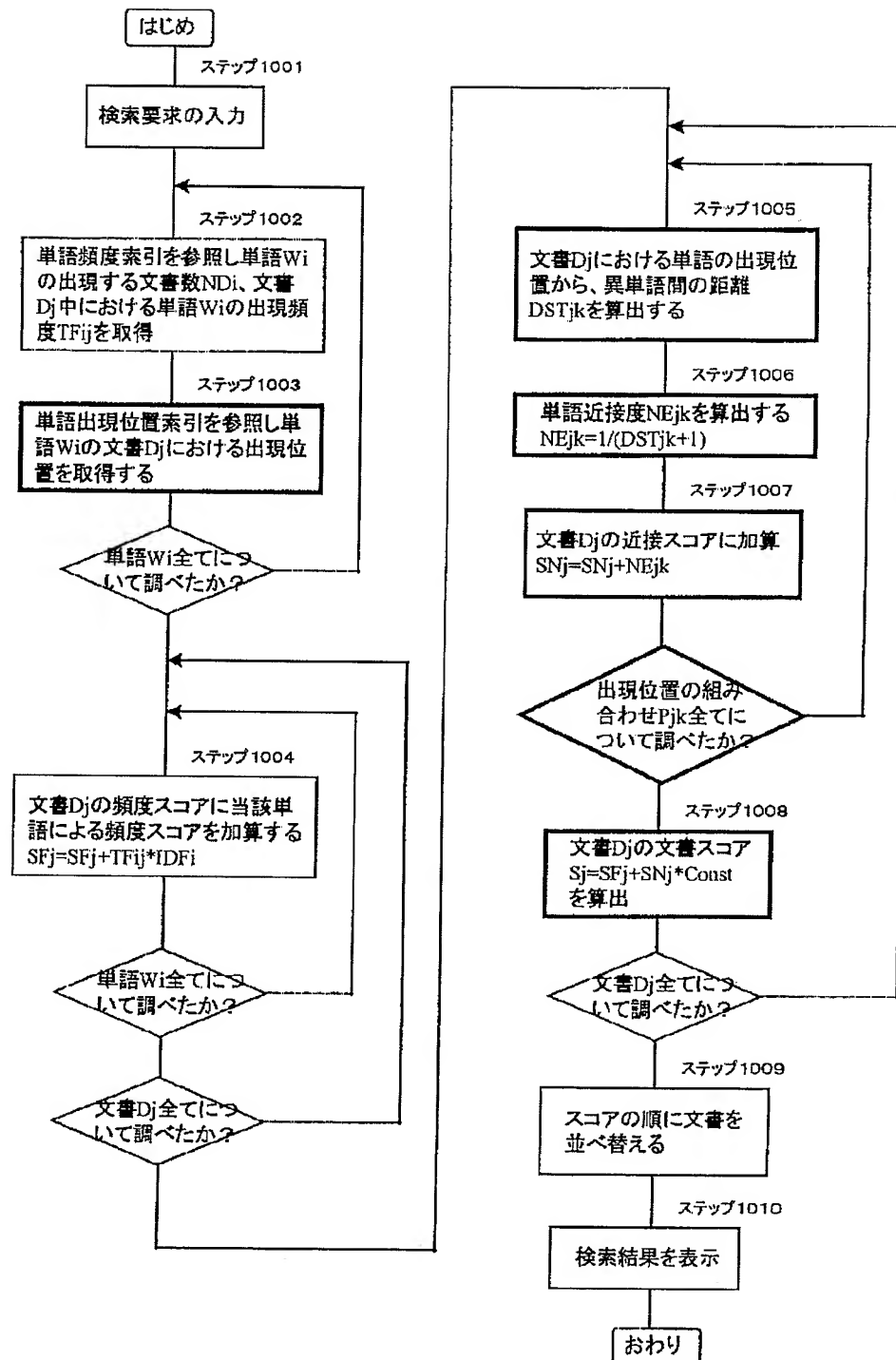
【図9】



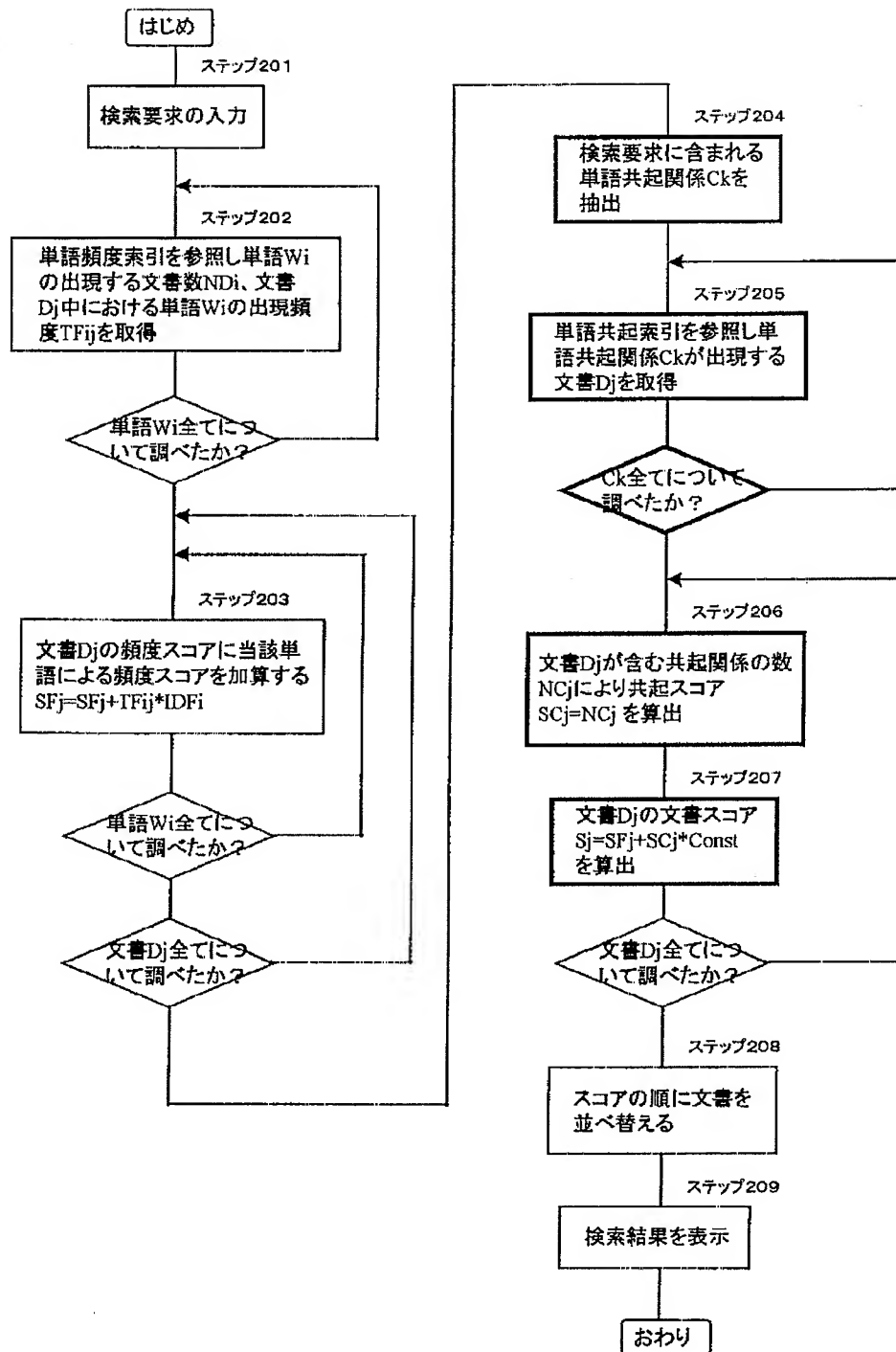
【図27】



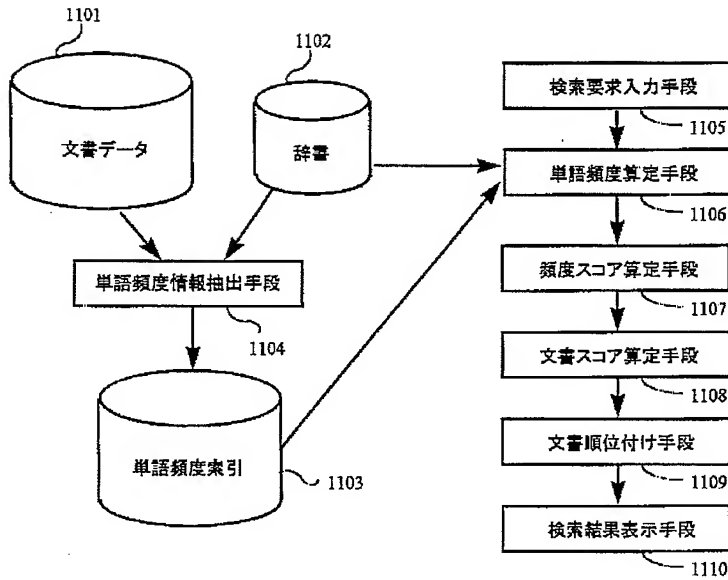
【図8】



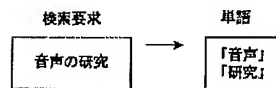
【図10】



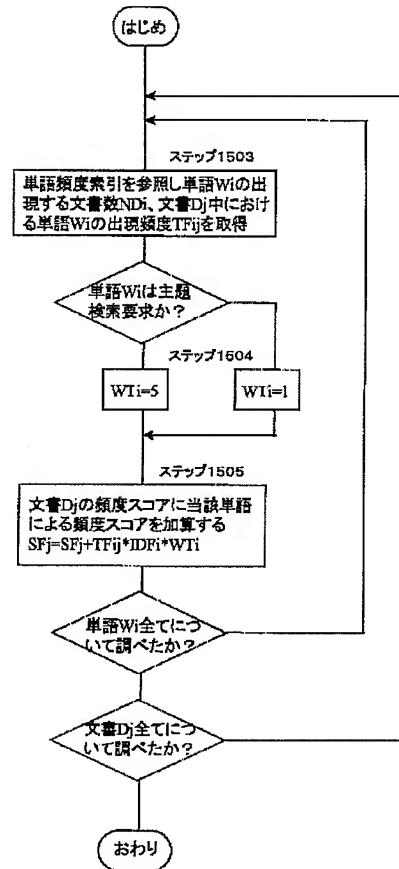
【図11】



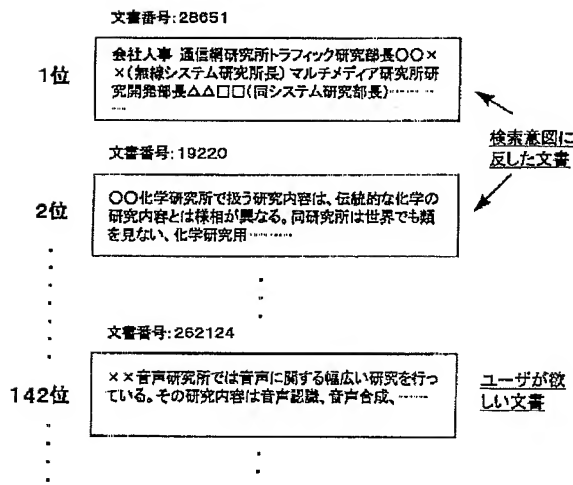
【図13】



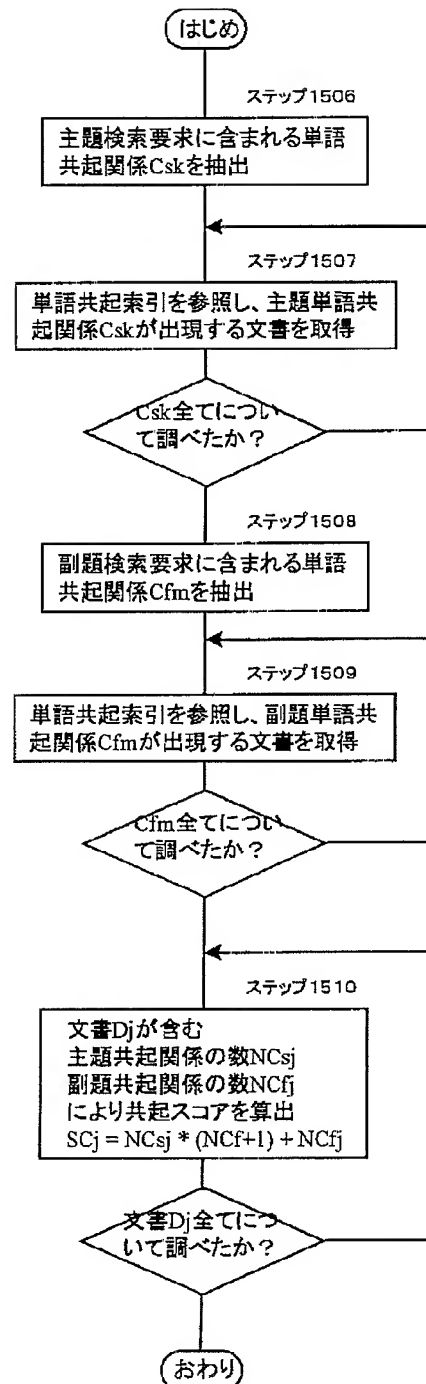
【図16】



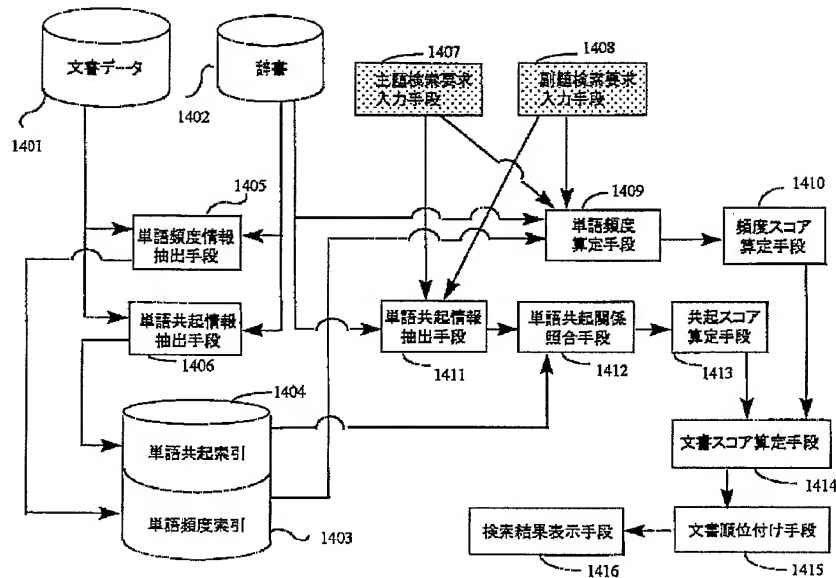
従来の頻度情報のみによる順位付け



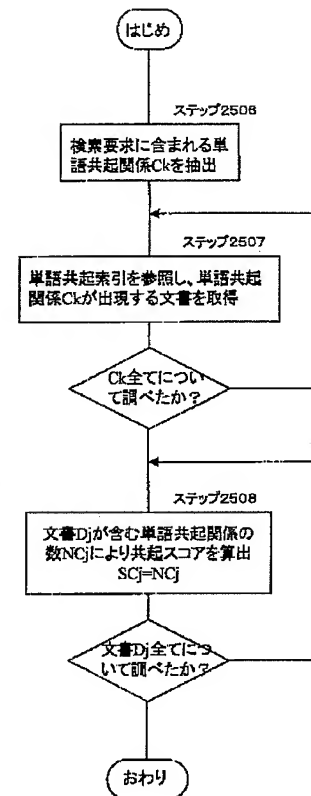
【图 17】



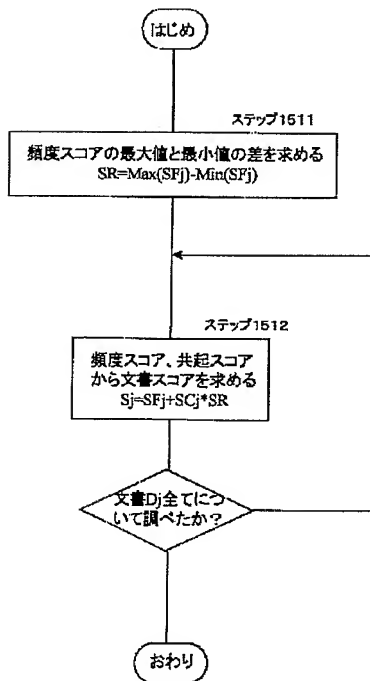
【図14】



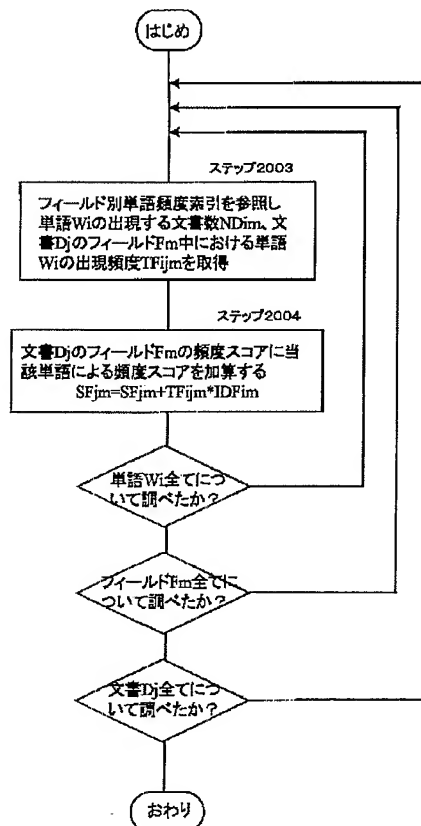
【図28】



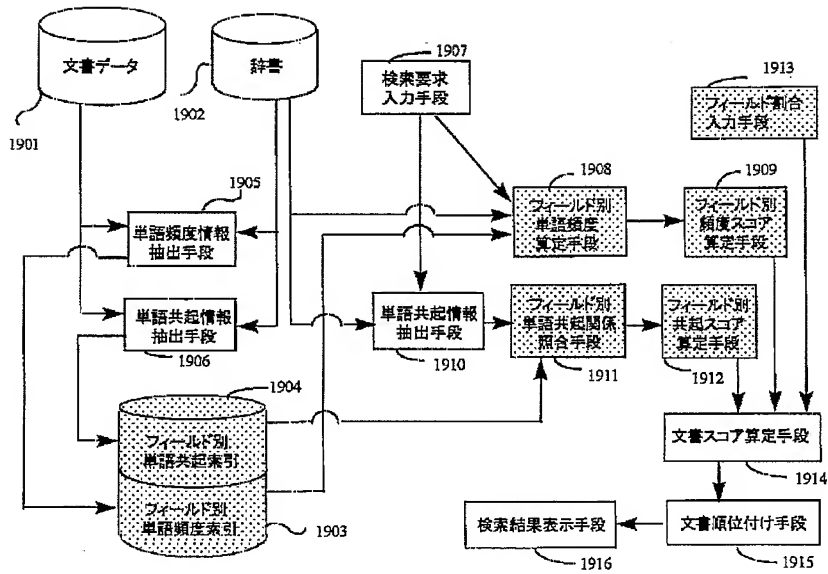
【図18】



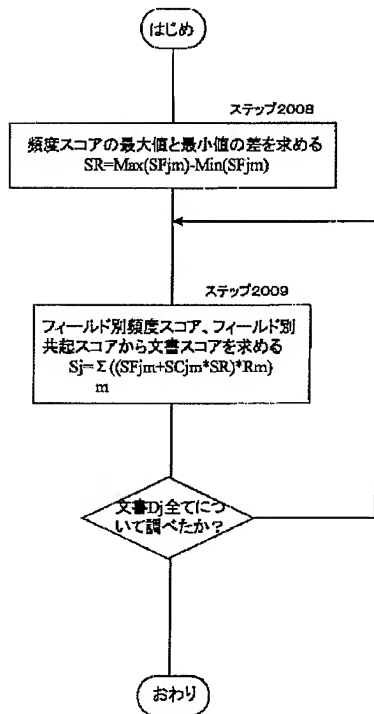
【図21】



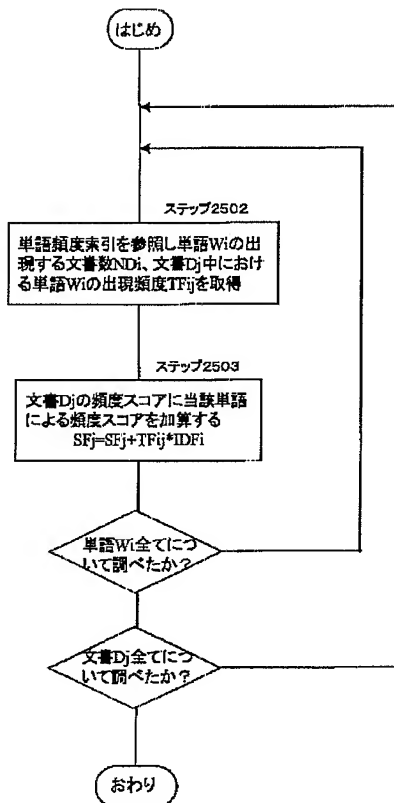
【図19】



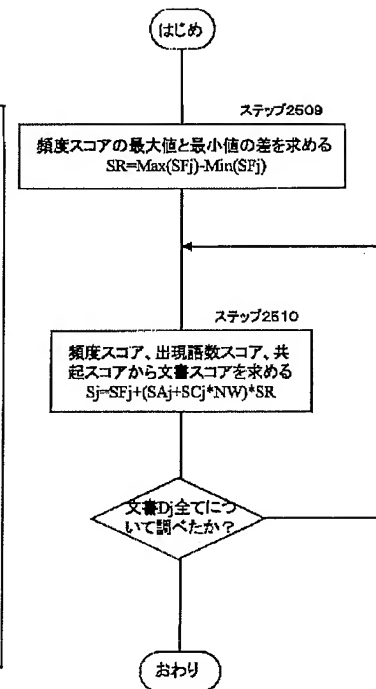
【図23】



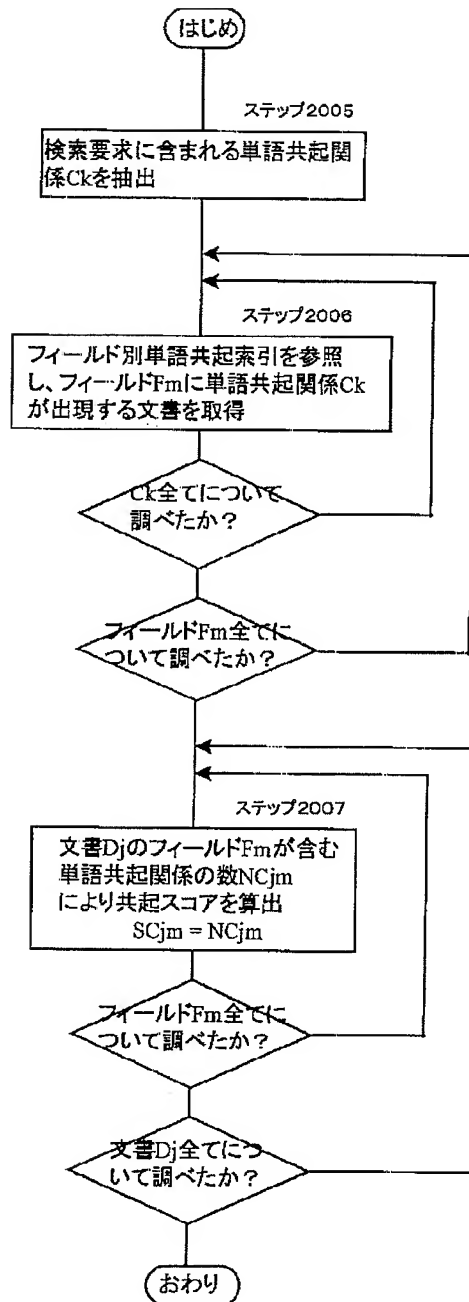
【図26】



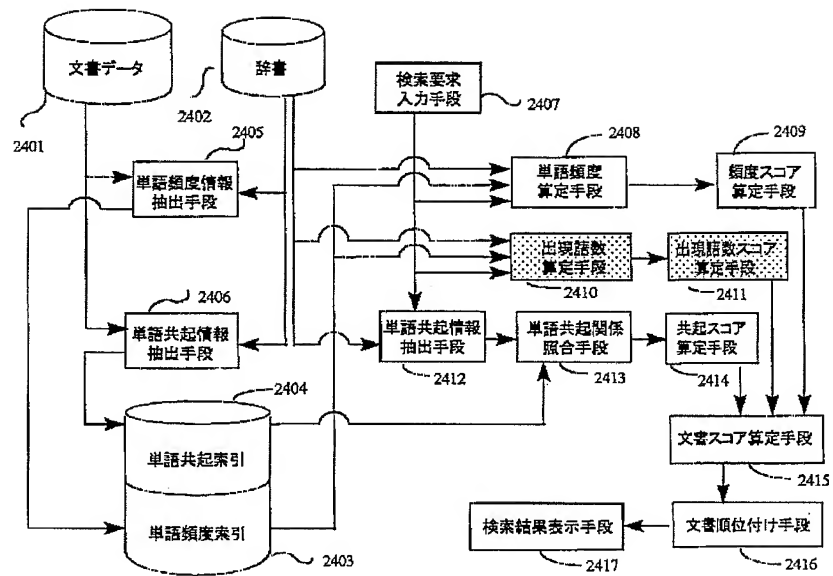
【図29】



【図22】



【図24】



フロントページの続き

(72)発明者 佐藤 光弘
大阪府門真市大字門真1006番地 松下電器
産業株式会社内

(72)発明者 野本 昌子
大阪府門真市大字門真1006番地 松下電器
産業株式会社内

(72)発明者 安川 秀樹
大阪府門真市大字門真1006番地 松下電器
産業株式会社内